<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>PAGE NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A New Approach for Adaptive Huffman Coding</td>
<td>1413</td>
</tr>
<tr>
<td>Pushpa Rani Suri, Madhu Goel</td>
<td></td>
</tr>
<tr>
<td>2. A New Variable Length Key Block Cipher Technique For Network</td>
<td>1423</td>
</tr>
<tr>
<td>Security In Data Transmission</td>
<td></td>
</tr>
<tr>
<td>A.V.N. Krishna, A.Vinaya Babu</td>
<td></td>
</tr>
<tr>
<td>3. A Comparative Study Of Clustering Algorithms For Building</td>
<td>1431</td>
</tr>
<tr>
<td>A Network Intrusion Detection Model</td>
<td></td>
</tr>
<tr>
<td>Mrutunjaya Panda, Manas Ranjan Patra</td>
<td></td>
</tr>
<tr>
<td>4. Semi Supervised Ensemble Clustering Algorithm (Ga Based)</td>
<td>1444</td>
</tr>
<tr>
<td>For High Dimensional Genomic Data</td>
<td></td>
</tr>
<tr>
<td>P.Krishnakumari, K.Vivekanandan</td>
<td></td>
</tr>
<tr>
<td>5. Combining Top Down Strategy With Bottom Up Approach</td>
<td>1455</td>
</tr>
<tr>
<td>For Image Segmentation</td>
<td></td>
</tr>
<tr>
<td>Mamata S. Kalas, P.P.Halkarnikar</td>
<td></td>
</tr>
<tr>
<td>6. An Improved Dwt Domain Statistical Image Watermarking System</td>
<td>1465</td>
</tr>
<tr>
<td>R. Brindha, S.C. Sharma</td>
<td></td>
</tr>
<tr>
<td>7. Integrating Heterogeneous Wireless Networks By Soft And Hard Handoff</td>
<td>1471</td>
</tr>
<tr>
<td>T.Sasikala, S.K.Srivatsa</td>
<td></td>
</tr>
<tr>
<td>8. Health Condition Observation With High Level Efficiency</td>
<td>1478</td>
</tr>
<tr>
<td>In An Advanced Sensor Network</td>
<td></td>
</tr>
<tr>
<td>Nagarajan M., Karthikeyan .S</td>
<td></td>
</tr>
<tr>
<td>9. Detection Of Myocardial Infarction Using Heart Rate Signals</td>
<td>1486</td>
</tr>
<tr>
<td>And Neural Networks</td>
<td></td>
</tr>
<tr>
<td>V. Mahesh, A.Kandaswamy</td>
<td></td>
</tr>
<tr>
<td>10. Complex Hadamard Matrices and Weighing Matrices</td>
<td>1492</td>
</tr>
<tr>
<td>M. K. Singh, Sudha Singh, D. K. Singh</td>
<td></td>
</tr>
</tbody>
</table>
A New Approach For Adaptive Huffman Coding

Pushpa Rani Suri1  Madhu Goel2

ABSTRACT

In this paper, the focus is on the use of ternary tree over binary tree in Huffman coding. First of all, we give the introduction of Huffman’s coding. Then adaptive Huffman coding is discussed. Here, a one pass Algorithm developed by Vitter for constructing adaptive Huffman codes using binary tree is implemented to ternary tree. In this paper, it is shown that the use of Ternary tree results in minimizing numbers of nodes (internal) and path length, fast implementation, efficient memory, fast compression ratio and error detecting & error correcting.

Keywords: Ternary tree, Huffman’s Algorithm, Adaptive Huffman coding, V algorithm, prefix codes, compression ratio, error detecting & correcting.

1. INTRODUCTION

Ternary tree or 3-ary tree is a tree in which each node has either 0 or 3 children (labeled as LEFT child, MID child, RIGHT child).

Data compression [15] or source coding is the process of encoding information using fewer bits or other information bearing units that an encoded representation would use through use of specific encoding schemes. As with any communication only works when both the sender and receiver understand that it is intended to be interpreted as characters representing the English language. Similarly compressed data can only be understood if the receiver knows the decoding method. Compression [1] is useful because it helps reduce the consumption of expensive resources such as hard disk or transmission bandwidth.

Huffman coding is an entropy-encoding algorithm used for loss less data compression. David A. Huffman [2] was given the concept of Huffman coding. Huffman coding uses a variable length [3] code table that has been derived in a particular way based on the estimated probability of occurrence for each possible value of the source symbol. Huffman coding uses a specific method for choosing the representation for each symbol, resulting in a prefix-free code that expresses the most common characters using shorter strings of bits than are used for less common source symbols. Huffman was able to design the most efficient compression method of this type: no other mapping of individual source symbols to unique string of bits will produce a smaller average output size when the actual frequencies agree with those used to create the code.

Huffman coding is divided in to two categories:-

1. Static Huffman coding
2. Adaptive Huffman coding

Static Huffman coding suffers from the fact that the uncompressed need have some knowledge of the probabilities of the symbol in the compressed files. This can need more bits to encode the file. If this information...
is unavailable compressing the file requires two passes. FIRST PASS finds the frequency of each symbol and constructs the Huffman tree. SECOND PASS is used to compress the file. We have already used the concept of static Huffman coding using ternary tree [10]. we conclude that representation of Static Huffman Tree using Ternary Tree is more beneficial than representation of Huffman Tree using Binary Tree in terms of number of internal nodes, Path length [13], height of the tree, in memory representation, in fast searching and in error detection & error correction. Static Huffman coding methods have several disadvantages.

Therefore we go for adaptive Huffman coding. Adaptive Huffman coding calculates the frequencies dynamically based on recent actual frequencies in the source string. Adaptive Huffman coding is also called dynamic Huffman coding. It is based on building the code as the symbols are being transmitted that allows one-pass encoding and adaptation to changing conditions in data. The benefits of one-pass procedure is that the source can be encoded in real time, through it becomes more sensitive to transmission errors, since just a single loss ruins the whole code.

Implementations of adaptive Huffman coding: -

There are number of implementations of this method, the most notable are

1. FGK (Faller Gallager Knuth) Algorithm
2. Vitter Algorithm

We have already used the concept of FGK Huffman coding using ternary tree [11]. We conclude that representation of FGK Huffman Tree using Ternary Tree is more beneficial than representation of Huffman Tree using Binary Tree in terms of number of internal nodes, Path length [12], height of the tree, in memory representation, in fast searching and in error detection & error correction.

Now here we try to use the concept of adaptive Huffman coding [4] using ternary tree with V Algorithm. All of these methods are defined- word schemes that determine the mapping from source messages to code- words on the basis of a running estimate of the source message probabilities. The code is adaptive, changing so as to remain optimal for the current estimates. In this way, the adaptive Huffman codes responds to locality, in essence, the encoder is learning the characteristics of the source. The decoder must learn along with the encoder by continually updating the Huffman tree so as to stay in synchronization with the encoder. Here we are given the concept of error detection and error correction. And the main point is that, this thing is only beneficial in TERNARY TREE neither in binary tree nor in other possible trees.

2. WHY WE USE ADAPTIVE HUFFMAN CODING

The key idea is to build a Huffman tree that is optimal for the part of the message already seen, and to recognize it when needed, to maintain its optimality. Adaptive Huffman [8] determines the mapping to code words using a running estimate of the source symbols probabilities.

1. It gives effective exploitation of locality. For example suppose a file starts out with the series of a character that are not repeated again in the file. In static Huffman coding that character will be low down on the tree because of its low overall count, thus taking lots of bits to encode. In adaptive Huffman coding, the character will be inserted at the highest leaf possible to be decoded, before eventually getting pushed down the tree by higher frequency characters.
2. Only one pass over the data.

3. Overhead, in static Huffman, we need to transmit someway the model used for compression that is the tree shape. This costs about $2n$ bits in a clever representation. As we will see, in adaptive schemes the overhead is $n \log n$.

3. CODING TECHNIQUE

3.1 Algorithm Vitter using Ternary Tree

V algorithm in Adaptive Huffman coding [7] uses binary tree, is extended to ternary tree.

In this section we discuss the one-pass algorithm V using ternary tree. The two main disadvantages of static Huffman’s algorithm are its two-pass nature and the overhead required to transmit the shape of the tree. In this paper we explore alternative one-pass methods, in which letters are encoded “on the fly”. We do not use a static code based on a single ternary tree, since we are not allowed an initial pass to determine the letter frequencies necessary for computing an optimal tree. Instead the coding is based on a dynamically varying Huffman tree. That is, the tree used to process the $t+1$ st letter is a Huffman tree with respect to $m_t$ the sender encodes the $t+1$ st letter $a_i$ in the message by sequence 00, 01 and 11 that specifies the path from root to leaf. The receiver then recovers the original letter by the corresponding traversal of its copy of the tree. Both sender and receiver then modify their copies of the tree before the next letter is processed so that it becomes a Huffman tree $\mu_{(t+1)}$.

$\mu_t = a_{t+1} \ a_t \ \ldots \ a_0$

The first $t$ letters in the message

The adaptive Huffman algorithm of Vitter (Algorithm V) incorporates two improvements over algorithm FGK.

It runs in real time and is optimum in our model of one-pass Huffman algorithms. There are two motivating factors in its design:-

1. The number of interchanges ($\uparrow$s) should be bounded by some small number. The number of interchanges in which a node is moved upward is limited to one. This number is bounded in algorithm FGK only by $L/2$ where $L$ is the length of the codeword for a $t+1$ when the recompilation begins.

2. The dynamic Huffman tree should be reconstructed to minimize not only $\sum w_j l_j$ but also $\sum l_j$ and max $\{l_j\}$, which intuitively has the effect of preventing a lengthy encoding of the next letter in the message.

Key Points Used in Vitter Algorithm

1. Swapping of nodes during encoding and decoding is onerous.

2. In FGK algorithm the number of swapping (considering a double cost for the updates that move a swapped node two levels higher) is bounded by $[dt/2]$, where $dt$ is the length of the added symbol in the old tree (this bound require some effort to be proved and is due to the work of Vitter)

3. In algorithm V, the number of swapping is bounded by 1.

These two objectives are reached through a new numbering scheme called implicit numbering.

Implicit Numbering

- One of the key ideas of Algorithm V is the use of a numbering scheme for the nodes that is different from the one used by algorithm FGK. We use an implicit numbering in which the node numbering corresponds to the visual representation of the tree that is the nodes of the tree are numbered in
increasing order by level; nodes on one level are numbered lower than the nodes on the next higher level.

- Nodes on the same level are numbered in increasing order from left to right.
- The node numbering used by algorithm FGK does not correspond to the implicit numbering

**Invariant**

- The key to minimize the other kind of interchanges is to maintain the following invariant
- For each weight w, all leaves of weight w precede (in the implicit numbering) all internal nodes of weight w.
- The interchanges, in the algorithm V are designed to restore implicit numbering, when a new symbol is read, and to preserve the invariant.

**Example 1**

Construct the V tree for the message (e eae deabe eae dcf) with ternary tree.

3.2 Algorithm V

1. The nodes of the tree are numbered in increasing order by level; nodes on one level are numbered lower than the nodes on the next higher level.

2. Nodes on the same level are numbered in increasing order from left to right.

3. If this numbering is satisfied, certain types of updates cannot occur.
3.3 Coding Technique For Ternary Tree

In Huffman Coding [12] the main work is to label the edges. Huffman Coding uses a specific method for choosing the representation for each symbol, resulting in a prefix-free code (sometimes called “Prefix Codes”) i.e. the bit string representing some particular symbol is never a prefix of the bit string representing any other symbol that expresses the most common characters using shorter strings of bits that are used for less common source symbols. The assignment entails labeling the edge from each parent to its left child with the digit 00, and the edge to the mid child with 01 and edge to the right child with 11. The code word for each source letter is the sequence of labels among the path from the root to the leaf node representing that letter. Only Huffman Coding is able to design efficient compression method of this type. Huffman Coding is such a widespread method for creating prefix-free codes that the term “Huffman Code” is widely used as synonym for “Prefix Free Code”. We will give a coding using variable length strings that is based on the Huffman Tree T for weighted data item as follows:

The Huffman Code [13] for Ternary Tree assigns to each external node the sequence of bits from the root to the node. Thus the above Tree T determines the code for the external nodes.
This code has “Prefix Property” i.e. the code of any item is not an initial sub string of the code of any other item. This means that there cannot be any ambiguity in decoding any message using a Huffman Code.

3.4 Compression Ratio (Fixed Length Code Cerses Huffman Length Code)

For example no. 1,

The number of fixed length code word bits= 4 bits (here in ternary tree, each symbol is represented by two bits, therefore for 7 symbols, number of fixed length code word bits are 4)

Average codeword length: -

\[ L_{ave} = \sum p_i l_i \]

\[ L_{ave} = 2 \times 0.38095 + 4 \times 0.147285 + 4 \times 0.23809 + 6 \times 0.04761 + 6 \times 0.09523 + 8 \times 0.04761 + 8 \times 0.04761 \]

\[ = 0.76194 + 5.891 + 9.523 + 28.56 + 57.13 + 38.08 + 38.08 \]

\[ = 3.9218 \]

In the above example, 7 symbols = 4 bits (fixed length code representation)

Lave (Huffman) = 3.9218 bits

Compression ration = 4/3.9218= 1.02

3.5 Error Detecting & Error Correcting

When this coding technique is applied in the message using ternary tree, then the number of transmitted bits is always even in number that is very beneficial in error detecting.

Error occurring during transmission is detected by following cases:

Case 1: Number of bits changed by addition or deletion of a bit.

Case 2: Prefix property is violated

Case 3: Sequence of bits does not exist as described in the labeling of edges in the coding technique.

If one of the cases occurs, accordingly can be corrected.

While In binary tree, the number of transmitted bits for a message can be either odd or even; therefore there is a difficulty in error detecting and in error correcting.

This thing is beneficial only in TERNARY TREE neither in binary tree nor in other possible trees.

3.6 Benefits Of Vitter Ternary Algorithm Over Vitter Binary Algorithm

Here, we are using message "e eae de eabe eae dcf" and then point out some comparison.
4. CONCLUSION

We can conclude that representation of Huffman Tree using Adaptive Ternary FGK Algorithm is more beneficial than representation of Huffman Tree using Adaptive Binary FGK Algorithm in terms of number of internal nodes, path length, height of the tree, in memory representation, in fast searching and in error detection & error correction.
REFERENCES


Author’s Biography

Dr. Pushpa Suri is a reader in the department of computer science and applications at Kurukshetra University Haryana India. She has supervised a number of PhD students. She has published a number of research papers in national and international journals and conference proceedings.

Mrs. Madhu Goel has Master’s degrees (University Topper) in Computer Science. At present, she is pursuing her PhD as University Research Scholar in Computer Science. Her area of research is Algorithms and Data Structure where she is working on Ternary search tree structures.
A New Variable Length Key Block Cipher Technique for Network Security in Data Transmission

A.V.N. Krishna, A. Vinaya Babu

ABSTRACT
The algorithm that is going to be discussed in this work will consider a random matrix key which on execution of sequence of steps generates a sequence. Based on the equality of values, this sequence is being divided into basins. The basins with minimum values will be eliminated. Remaining each basin represents one block of data. Depending on starting input value of plain text character, corresponding basin is considered as a key. The procedure is repeated for certain plain text depending on chosen value. Thus the cipher text obtained becomes very difficult to be broken with out knowing the key.

Keywords: Cryptography, Variable length key, Encryption Algorithm, Example, Add function.

1. INTRODUCTION
A crypto system [ref 1-5] is an algorithm, plus all possible plain texts, cipher texts and keys. There are two general types of key based algorithms: symmetric and public key. With symmetric-key encryption, the encryption key can be calculated from the decryption key and vice versa. With most symmetric algorithms, the same key is used for both encryption and decryption, as shown in Figure 1.

Implementations of symmetric-key encryption can be highly efficient, so that users do not experience any significant time delay as a result of the encryption and decryption. Symmetric-key encryption also provides a degree of authentication, since information encrypted with one symmetric key cannot be decrypted with any other symmetric key. Thus, as long as the symmetric key is kept secret by the two parties using it to encrypt communications, each party can be sure that it is communicating with the other as long as the decrypted messages continue to make sense.

Symmetric-key encryption is effective only if the symmetric key is kept secret by the two parties involved. If anyone else discovers the key, it affects both confidentiality and authentication. A person with an unauthorized symmetric key not only can decrypt messages sent with that key, but can encrypt new messages and send them as if they came from one of the two parties who were originally using the key.

2. A NEW ALGORITHM [6-12 &18-19]
2.1 The New Algorithm Has the Following Features
1. A set of poly alphabetic substitution rule is used.
2. A Random matrix Key is used as a key.
3. A sequence is generated from a product of ternary vector and the key.

4. Based on equality of values, the sequence is divided into basins.

5. These basins are used as variable length keys to the characters in the plain text by a particular chosen rule.

2.2 The New Algorithm Is Combination Of

a. Substitution cipher

b. A Random Matrix key
c. Development of basins.
d. Each basin with unequal number of values.
e. Modified Caesar algorithm.
f. Coding method.

Thus this algorithm is a combination of

1. A Random vector of 27 values to few basin values.

2. Each basin containing different number of values.

3. Corresponding basin is considered as key depending on first character of plain text.

2.3 The Steps That Are Involved In the Proposed Algorithm

1. The letters of the alphabet were given numerical values starting from 0

2. A random matrix used as a key. Let it be A.

3. A ternary vector of 27 values is considered.

4. Ternary vector is multiplied with Random matrix key.

5. A Sign function is applied on the product to make it more secure.[14-17]

6. A sequence is generated.

7. Basins will be developed from this sequence which contains similar values.

8. Thus each basin contains unequal number of values.

9. Basins with minimum values (say 1 value) are eliminated.

10. Each basin is represented by one character.

11. The plain text being converted to equivalent numerical value.

12. The alphanumerical value of the first character of plain text is considered.

13. The mod of the first value with the number of basins is calculated.

14. Depending on the value of the remainder, the corresponding basin is considered as a key.

15. The key is added to the plain text to generate cipher text.

16. The procedure is repeated for successive plain texts depending on chosen value.

3. Advantages

1. It is almost impossible to extract the original information.

2. Even if the algorithm is known, it is difficult to extract the information.

3. Versatile to users. Different users of internet can use different modified versions of the new algorithm.

4. As per basin values, the same character is substituted by different alpha numerical value which provides more security for the message.
A New Variable Length Key Block Cipher Technique for Network Security in Data Transmission

4. Example: \( n = 0 : 26 \);
\[ r = \text{ternary vector } 0:26 \]
\[ n = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 26 \]
\[ r = 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 2 \]
\[ r = r-1; \]
\[ r = -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ -1 \quad -1 \quad -1 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ A = \text{key} = \begin{bmatrix} 2 & 5 & -6 \\ 3 & 1 & 3 \\ 4 & -2 & -3 \end{bmatrix} \]
\[ r = \text{multiply } (A*r). \]
\[ r = -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ -1 \quad -1 \quad -1 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ r = r+1; \]
\[ r = 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 2 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 2 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]

Thus \( r = r+(2,1)*3+r(1,1)*9; \)
\[ r=r(3,1)+r(2,1)*3+r(1,1)*9; \]
\[ \text{thus for } N = 0; \quad r = (2 \quad 0 \quad 0) = 2 + 0 * 3 + 0 * 9 = 2; \]

Ternary vector \( n=0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 17 \quad 18 \quad 19 \quad 20 \quad 21 \quad 22 \quad 23 \quad 24 \quad 25 \quad 26 \)

Random Sequence generated

Thus \( r = 2 \quad 0 \quad 0 \quad 18 \quad 0 \quad 3 \quad 18 \quad 18 \quad 6 \quad 20 \quad 6 \quad 20 \quad 13 \quad 6 \quad 20 \quad 24 \quad 6 \quad 20 \quad 8 \quad 8 \quad 23 \quad 26 \quad 8 \quad 26 \quad 26 \quad 24. \)

The basins that can be formed using this sequence are
\[ b(0) = (0,1,2,4,10) \]
\[ b(1) = (3,5,6,7,8,9,11,12,14,15,17,18,19,20,23) \]
\[ b(2) = (16,21,22,24,25,26) \]

Plain text:avnkrishna

Considering the first character of the plain text, \( a = 10, \)
\( 10 \mod 3 = 1. \) So the basin considered is \( b(1) \)

1. Encryption
1. Plain text
\[ a \quad v \quad n \quad k \quad r \quad I \quad s \quad h \quad n \quad a \]
\[ 10 \quad 31 \quad 23 \quad 20 \quad 27 \quad 18 \quad 28 \quad 17 \quad 23 \quad 10 \]
2. Basin Considered. \( b(1) \)
\[ 3 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 11 \quad 12 \quad 14 \quad 15 \]
3. Add Function
\[ 13 \quad 36 \quad 29 \quad 27 \quad 35 \quad 27 \quad 39 \quad 29 \quad 37 \quad 25 \]
4. Mod 36
\[ 13 \quad 0 \quad 29 \quad 27 \quad 35 \quad 27 \quad 10 \quad 29 \quad 10 \quad 25 \]
5. Cipher text
d0t z r a t a p

2. Decryption
1. Cipher text
d0t z r a t a p
13 0 29 35 27 10 29 10 25
2. Mod 36
13 36 29 35 27 39 29 37 25
3. Basin Considered. b(1)
3 5 6 7 8 9 11 12 14 15
4. Subtract function
10 31 23 27 18 28 17 23 10
5. Plain text
a v n k r l s h n a.
After some part of plain text, the procedure is repeated.

5. Total Number Of Computations
Considered in the given model for converting plain text to cipher text.

1 computation: converting n=0:26 to ternary vector.
   Let it be r.
2 computation: calculating r-1.
3 computation: multiplying r with the key considered.
4 computation: applying sign function on the product.
   Store it in r.
5 computation: calculating r+1
6 computation: converting output ternary vector to integer form. Let this be s, the sequence generated.
7 computation: Different basins will be formed by placing similar values of the sequence in one basin.
8 computation: converting plain text to alphanumerical value.
9 computation: considering the first character of plain text and applying a mod function of the order of number of basins formed.
10 computation: depending on the output of the mod function, the corresponding basin be used as a key.
11 computation: add the key to the alphanumerical value of the plain text.
12 computation: applying mod function on the output.
13 computation: converting the output to characters of the alphabet to get cipher text.
14 computation: The procedure being repeated for chosen part of plain text.
Thus the total number of computations in the first proposed model are 15.

Computation overhead for a 27 character key.

1 computation: 27 calculations.
2 computation: 27 calculations
Key considered: 9 character key.
3 computation: 27*9 calculations.
4 computation: 27 calculations.
5 computation: 27 calculations
6 computation: 27 calculations
7 computation: 27*27 calculations + 27*27 calculations
8 computation: 27 calculations.
9 computation: 03 calculations depending on chosen rule.
10 computation: 03 calculations.
11 computation: 27 calculations
12 computation: 27 calculations
13 computation: 27 calculations
A New Variable Length Key Block Cipher Technique for Network Security in Data Transmission

14 computation: 03 calculations depending on the chosen rule.

Thus the total computational overhead by the proposed model is 1953 calculations.

6. Security Analysis by Construction

6.1 If We Go By the Construction of the Algorithm

1. computation: converting n=0:26 to ternary vector. Let it be r. The complexity is multiples of n.

2. computation: calculating r-1. The complexity is multiples of n.

3. computation: multiplying r with the key considered. The complexity is multiples of n.

4. computation: applying sign function on the product. Store it in r. The complexity is multiples of n.

5. computation: calculating r+1. The complexity is multiples of n.

6. computation: converting output ternary vector to integer form. Let this be s, the sequence generated. The complexity is multiples of n.

7. computation: 27*27 calculations + 27*27 calculations. The complexity is multiples of n*n.

8. computation: 27 calculations. The complexity is multiple of n.

9. computation: 03 calculations depending on the chosen rule. The complexity depends on chosen rule.

10. computation 03 calculations. The complexity depends on chosen rule.

11. computation: 27 calculations. The complexity is multiple of n.

12. computation: 27 calculations. The complexity is multiple of n.

13. computation: 27 calculations. The complexity is multiple of n.

14. computation: 03 calculations depending on the chosen rule.

The complexity depends on chosen rule.

Thus the total computational complexity of the model by its construction is of the order of O(n^2).

6.2 Complexity of the Algorithm by Its Strength

In the given algorithm a matrix key is used. This matrix key is multiplied with ternary vector. On the generated values a sign function is used to convert all positive values to 1, negative values to -1 and zero to 0. This provides the necessary strength to the algorithm. Thus known the algorithm, known the plain text & cipher text pairs, it is computationally infeasible to generate the matrix key. Thus this algorithm is supposed to be safe in real time environment. Thus if we go by the security of the algorithm, the complexity is exponential in nature.

7. Avalanche Effect

In this model a sequence is generated and this sequence is used to generate basins of variable length. These basins will be used as keys for the plain text to generate the cipher text. Depending on the key; the sequence will be used to generate the basins. We will identify the variations in the basins generated, by slight variations in the key. Thus we can identify the variations in the cipher text by slight variations in the key considered. We will also identify the variations in the cipher text by slight variations in the plain text. For example,

Case 1:

\[
A=\begin{pmatrix}
2 & 5 & -6 \\
3 & 1 & 3 \\
4 & -2 & -3 
\end{pmatrix}
\]

Sequence generated from the proposed model:

n=0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
r=2 0 0 0 0 0 0 3 18 18 6 20 2 6 20 13 6 20 24 6 20
Thus the basins that can be formed using these sequences which are used as key are:

\[ b(0) = (0,1,2,4,10) \]
\[ b(1) = (3,5,6,7,8,9,11,12,14,15,17,18,19,20,21,23) \]
\[ b(2) = (13) \]
\[ b(3) = (16,22,24,25,26) \]

**Case 2:** By increasing the values of key by 1 at each row.

\[ A = \text{key} = \begin{array}{cccc}
3 & 5 & -6 \\
4 & 1 & 3 \\
5 & -2 & -3 
\end{array} \]

Sequence generated:
\[ r = 1 \quad 0 \quad 0 \quad 18 \quad 0 \quad 0 \quad 18 \quad 18 \quad 3 \quad 20 \quad 2 \quad 6 \quad 20 \quad 13 \quad 6 \quad 20 \quad 24 \quad 6 \quad 23 \quad 8 \quad 8 \quad 26 \quad 26 \quad 8 \quad 26 \quad 26 \quad 25. \]

Thus the basins formed which are used as key are:

\[ b(0) = (0,1,2,4,5,10) \]
\[ b(1) = (3,6,7,8,9,11,12,14,15,17,18,19,20,23) \]
\[ b(2) = (13) \]
\[ b(3) = (16,21,22,24,25,26) \]

**Case 3:** By decreasing the key values by 1 at each row.

\[ A = \text{key} = \begin{array}{cccc}
1 & 5 & -6 \\
2 & 1 & 3 \\
2 & -2 & -3 
\end{array} \]

Sequence generated by the model:
\[ r = 11, 1, 3, 20, 0, 6, 18, 18, 6, 20, 2, 6, 20, 13, 6, 20, 24, 6, 20, 8, 20, 26, 6, 23, 25, 15. \]

Basins formed which are used as key are:

\[ b(0) = (0,11,4) \]
\[ b(1) = (1) \]
\[ b(2) = (2,3,10) \]
\[ b(3) = (5,6,7,8,9,11,12,14,15,17,18,19,20,21,22,23,24,26) \]
\[ b(4) = (13) \]
\[ b(5) = (25). \]

**For example**

By converting the characters of the alphabet to alphanumerical values, starting from 1,

- If the plain text considered is \( \text{a b c d e} \)
  - Its equivalent alphanumerical value is 01 02 03 04 05.
  - As per the algorithm key is \( b(0) = 01 03 05 08 15 \)
  - The cipher text formed will be \( \text{a c e h o} \).

If the key is decreased by 1 at each row,

- If the plain text considered is \( \text{a b c d e} \)
  - Its equivalent alphanumerical value is 1 2 3 4 5.
  - As per the algorithm key is \( b(0) = 01 12 04 01 11 \)
  - The cipher text formed will be \( \text{a m g d p} \).

Thus we can see that, by changing the key slightly, there are a lot of variations in the basins generated. Since these basins will form the keys for the given model, they provide maximum avalanche effect to the algorithm. Thus this model provides for good variations in cipher text for slight variations in the key. This provides for maximum strength and security to the algorithm.

If the plain text is slightly varied, a different basin will be considered as key depending on the first value of plain text.

**For example**

If the plain text considered is \( \text{a b c d e} \)

- Its equivalent alphanumerical value is 01 02 03 04 05.
- As per the algorithm key is \( b(0) = 01 02 03 04 05 \)
- The cipher text formed will be \( \text{a c e h o} \).

If the plain text is slightly varied by one value,

- If the plain text considered is \( \text{b b c d e} \)
  - Its equivalent alphanumerical value is 02 02 03 04 05.
  - As per the algorithm key is \( b(1) = 3 5 6 7 8 \)
A New Variable Length Key Block Cipher Technique for Network Security in Data Transmission

The cipher text formed will be 5 7 9 11 13 e g I k m.

We will see a lot of variations in the cipher text generated for slight variations in the plain text. Thus it provides a maximum avalanche effect to the algorithm which provides for more strength and security.

8. SECURITY ANALYSIS

1. A Variable length key is used which makes the algorithm free from linear cryptanalysis.
2. A sign function is used for development of sequence which is used to generate basins. This sign function converts all positive values to 1, negative values to -1 & zero to 0. This makes the algorithm free from differential crypto analysis.
3. A mod function is used on selective plain text values to choose corresponding basin. This provides for better avalanche effect with a small change in plain text will produce a significant variations in cipher text.
4. A small change in key values will have a considerable change on the values of basins, which provides for better avalanche effect.

9. COMPARATIVE STUDY OF THE PROPOSED VARIABLE LENGTH KEY ENCRYPTION ALGORITHM WITH STANDARD ALGORITHMS LIKE DES AND RSA IN TERMS OF COMPUTATIONAL OVERHEAD, COMPLEXITY AND SECURITY ANALYSIS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Computational overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>DES</td>
<td>64/24 for a 56 bit key</td>
</tr>
<tr>
<td>RSA</td>
<td>1,26,072 for a 200 bit key</td>
</tr>
<tr>
<td>A new variable length key</td>
<td>1953 for a 9 character</td>
</tr>
</tbody>
</table>

10. CONCLUSION

The present work is compared with standard algorithms like DES & RSA. It is observed that the proposed algorithm is having equal security with respect to DES & RSA algorithms at low computational overhead. Thus in applications like Broad casting applications, wireless sensor networks the proposed algorithm can well be used. The proposed work along with public key distribution mechanism can well be used in secure electronic transactions, digital cash where security to data is of higher importance. It is also observed that by changing the key by one value, the number of basins formed is varying in nature, which provides a better avalanche effect. This provides more security and strength to the algorithm.

REFERENCES


Author’s Biography

A.V.N.Krishna, Professor, Computer Science, Indur Institute of eng. & Tech., Siddipet, India. He has 19 yrs of teaching Experience at UG & PG level and actively involved in Research. His fields of interest are Cryptography, Mathematical modeling & Data Mining. He has Published Papers at Journals of National & International repute and presented work at International Conferences.
A Comparative Study of Clustering Algorithms for Building a Network Intrusion Detection Model

Mrutyunjaya Panda1 Manas Ranjan Patra2

ABSTRACT

K-means is a popular clustering algorithm that requires a huge initial set to start the clustering. K-means is an unsupervised clustering method which does not guarantee convergence. Numerous improvements to K-means have been done to make its performance better. Now fuzzy set theory has been applied to many fields including data mining. Fuzzy clustering method is more precise in dealing with data simulation, and the results are easier to be understood and used. Therefore, research into fuzzy clustering method for knowledge is significant not only to theory, but also to application. Expectation Maximization is a statistical technique for maximum likelihood estimation using mixture models. It searches for a local maxima and generally converges very well. In this paper, we propose some clustering algorithms such as K-Means, Fuzzy c-Means, and EM (Expectation and maximization) for network intrusion detection. We have used KDDCup’1999 data set for our experimentation. The simulation results show that EM algorithm is a more statistically formalized method, which accounts for partial membership in classes. It has better convergence properties and is in general preferred to K-Means and Fuzzy c-Means algorithms in building a network intrusion detection model.

Keywords: Intrusion Detection, K-Means, Fuzzy c-Means, EM, MF plot, ROC, log likelihood.

1. INTRODUCTION

An intrusion detection system (IDS) is a component of the information security framework. Its main goal is to differentiate between normal activities of the system and behaviour that can be classified as suspicious or intrusive [1]. The goal of intrusion detection is to build a system which would automatically scan network activity and detect such intrusion attacks. Once an attack is detected, the system administrator can be informed who can take appropriate action to deal with the intrusion. IDS can be host-based (HIDS), network-based (NIDS) or a combination of both types (Hybrid Intrusion Detection System). HIDS usually observes logs or system calls on a single host, while a NIDS typically monitors traffic flows and network packets on a network segment, and thus observes multiple hosts simultaneously. Generally, one deal with very large volumes of network data, and thus it is difficult and tiresome to classify them manually in order to detect a possible intrusion. One can obtain labelled data by simulating intrusions, but this will be limited only to the set of known attacks. Therefore, new types of attacks that may occur in future cannot be handled, if those were not part of the training data. Even with manual classification, we are still limited to identifying only the known (at classification time) types of attacks, thus restricting our detection system to identifying only those types.

1Dept. of ECE, GIET, Gunupur, Orissa, India. E-mail: mrutyunjaya.2007@rediffmail.com
2Dept. of Computer Science, Berhampur University, Orissa, India. E-mail: mrpatra12@gmail.com
To solve these difficulties, we need a technique for detecting intrusions when our training data is unlabeled, as well as for detecting new and un-known types of intrusions. A method that offers promise in this task is anomaly detection. Anomaly detection detects anomalies in the data (i.e. data instances in the data that deviate from normal or regular ones). It also allows us to detect new types of intrusions, because these new types will, by assumption, be deviations from the normal network usage.

It is very difficult, if not impossible, to detect malicious intent of someone who is authorized to use the network and who uses it in a seemingly legitimate way. For example, there is probably no highly reliable way to know whether someone who correctly logged into a system is the intended user of that system, or if the password was stolen.

Under these assumptions we built a system which created clusters from its input data, then automatically labelled clusters as containing either normal or anomalous data instances, and finally used these clusters to classify network data instances as either normal or anomalous. Both the training and testing was done using 10% KDDCup’99 data [2], which is a very popular and widely used intrusion attack dataset.

A popular algorithm is the K-means where, based on a given number of clusters, the algorithm iterates to find best clusters for the objects. Most clustering techniques assume a well defined distinction between the clusters so that each pattern can only belong to one cluster at a time. This supposition can neglect the natural ability of objects existing in multiple clusters. For this reason and with the aid of fuzzy logic, fuzzy clustering can be employed to overcome the weakness. The membership of a pattern in a given cluster can vary between 0 and 1. In this model a data object belongs to the cluster where it has the highest membership value. In this paper, we aim to propose a fuzzy c-means clustering technique which is capable of clustering the most appropriate number of clusters based on objective function. This, as the name implies, draws the fuzzy boundary, thereby proving efficient when compared with that of its counterpart. Another approach is to use the Expectation Maximization algorithm (EM). The Expectation Maximization algorithm may be a very efficient technique for estimating class conditional probability density Functions (PDF) in both univariate and multivariate cases [3]. This paper discusses about the K-Means, Fuzzy c-Means and EM clustering algorithms and compares their suitability in building an efficient network intrusion detection model.

The rest of the paper is organised as follows. In section 2, we discuss Clustering Methods; followed by Data Clustering algorithms in section 3. In Section4, various cluster evaluation schemes are discussed. Section 5 describes about the experimental set-up and results obtained. Some discussion is made in Section 6. Finally, Section 7 provides some related works followed by conclusion in Section 8.

2. CLUSTERING METHODS

Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory) [4]. By definition, “cluster analysis is the art of finding groups in data”, or from Wikipedia [5], “clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data into subsets (clusters), so that the data in each subset (ideally) share some common trait-often proximity
according to some defined distance measure. Clustering is a challenging field of research as it can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, cluster analysis serves as a pre-processing step for other algorithms, such as classification which would then operate on detected clusters.

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. By clustering, one can identify dense and sparse regions and therefore, discover overall distribution patterns and interesting correlations among data attributes. Clustering does not rely on predefined classes and class labelled training examples. For this reason, clustering is a form of learning by observation.

In intrusion detection, an object is a single observation of audit data and/or network packets after the values from selected features have been extracted. Hence, values from selected features, and one observation, define one object (or vector). If we have values from n number of features, the vector (or object plot) fits into an n-dimensional coordinate system (Euclidean space R^n).

In order to derive the objective function and other relevant mathematics for fuzzy c-means and the remaining of its variations, it is better to see the same for the hard (crisp) partitioning technique, so that we may be able to understand the difference between the two approaches. (If we look into these issues all of them appears to be objective functional minimization problems. If the constraints are relaxed we get the possibilistic partition scheme. So, the clustering algorithm is nothing but a minimization problem which may be constrained or unconstrained.)

2.1. Hard Partitioning

These kind of methods are based on classical set theory and defines the presence or absence of a data point in a partition subset on strict logic, that is the object either belong to a subset or not. So, such kind of methods divides a dataset strictly into disjoint subsets.

Conventional clustering algorithms find a “hard partition” of a given data set based on certain criteria that evaluate the goodness of a partition. By hard partition we mean that each datum belongs to exactly one cluster of the partition. More formally, we can define the concept of “hard partition” as follows:

1) Let X be a data set of data, and x_i be an element of X. A partition P={C_1, C_2,…,C_j} of X is “hard” if and only if
   i) \( \forall x_i \in X \exists C_j \in P \text{ such that } x_i \in C_j \)
   ii) \( \forall x_i \in X, x_i \in C_j \Rightarrow x_i \notin C_k \text{ where } k \neq j, C_j \in P. \)

2.2. Soft Partitioning

A soft clustering algorithm partitions a given data set not an input space. Theoretically speaking, a soft partition not necessarily a fuzzy partition, since the input space can be larger than the dataset. In practice however most
soft clustering algorithms do generate a soft partition that also forms the fuzzy partition.

A type of soft clustering of special interest is one that ensures the membership degree of a point \( x \) in all clusters adding up to one, i.e.

A soft partition that satisfies this additional condition is called a constrained soft partition. The fuzzy c-means algorithm produces a constrained soft partition. The fuzzy c-means algorithm is best known algorithm that produces constrained soft partition.

The biggest drawback of a hard partitioning is the concept that it either includes a data point in a partition or strictly excludes it; there is no other chance for the data elements to be part of more than one partition at the same time. However, in natural clusters it is always the case that some of the data elements partially belong to one set and partially to one or more other sets. In order to overcome this limitation, the notion of fuzzy partitioning was introduced [6].

3. DATA CLUSTERING ALGORITHMS

The following are the algorithms used for clustering the datasets:

- K-means Algorithm
- Fuzzy c-means Algorithm.
- Expectation-Maximization (EM) Algorithm.

3.1. K-means Clustering

The k-means clustering is a classical clustering algorithm. After an initial random assignment of example to \( k \) clusters, the centres of clusters are computed and the examples are assigned to the clusters with the closest centres. The process is repeated until the cluster centres do not significantly change. Once the cluster assignment is fixed, the mean distance of an example to cluster centres is used as the score. Using the K-means clustering algorithm, different clusters were specified and generated for each output class [7].

K-means clustering is a well known Data Mining algorithm that has been used in an attempt to detect anomalous user behaviour, as well as unusual behaviour in network traffic. There are two problems that are inherent to k-means clustering algorithms. The first is determining the initial partition and the second is determining the optimal number of clusters [8]. In figure 1 depicted K-means algorithm.

**K-MEANS ALGORITHM**

Input: The number of clusters \( K \) and a dataset for intrusion detection

Output: A set of \( K \)-clusters that minimizes the squared error criterion.

Algorithm:

1. Initialize \( K \) clusters (randomly select \( K \) elements from the data)
2. While cluster structure changes, repeat from 2.
3. Determine the cluster to which source data belongs
   Use Euclidean distance formula. Add element to cluster with min (distance \( (x_i, y_j) \)).
4. Calculate the means of the clusters.
5. Change cluster centroids to means obtained using Step 3.

Figure 1: K-means Clustering

As the algorithm iterates through the training data, each cluster’s architecture is updated. In updating clusters, elements are removed from one cluster to another. The updating of clusters cause the values of the centroids to change, which is a reflection of the current cluster elements. Once there are no changes to any cluster, the training of the K-Means algorithm is complete.

At the end of the K-Means training, the K cluster centroids are created and the algorithm is ready for
A Comparative Study of Clustering Algorithms for Building a Network Intrusion Detection Model

classifying traffic. For each element to be clustered, the cluster centroids with the minimal Euclidean distance from the element will be the cluster for which the element will be a member. After training, the cluster centroids remains the same, like the SOM (Self organise Map) can be useful for anomaly detection tool that requires the input to remain static. The k-Means algorithm may take a large number of iterations through dense data sets before it can converge to produce the optimal set of centroids. This can be inefficient on large data sets due to its unbounded convergence of cluster centroids.

3.2. Fuzzy c-Means (FCM) Clustering

Fuzzy c-Means (FCM) algorithm, also known as fuzzy ISODATA, was introduced by Bezdek [9] as extension to Dunn’s [10] algorithm to generate fuzzy sets for every observed feature. The fuzzy c-means clustering algorithm is based on the minimization of an objective function called c-means functional.

Fuzzy c-means algorithm is one of the well known relational clustering algorithms. It partitions the sample data for each explanatory (input) variable into a number of clusters. These clusters have “fuzzy” boundaries, in the sense that each data value belongs to each cluster to some degree or other. Membership is not certain, or “crisp”. Having decided upon the number of such clusters to be used, some procedure is then needed to location their centres (or more generally, mid-points) and to determine the associated membership functions and the degree of membership for the data points.

Fuzzy clustering methods allow for uncertainty in the cluster assignments. FCM is an iterative algorithm to find cluster centres (centroids) that minimize a dissimilarity function. Rather that partitioning the data into a collection of distinct sets by fuzzy partitioning, the membership matrix (U) is randomly initialized according to equation 2.

\[ \sum_{i=1}^{c} u_{ij} = 1, \forall j=1,2,\ldots,n. \]  

(2)

The dissimilarity function (or more generally the objective function), which is used in FCM in given equation 3.

\[ J(U,c_1,c_2,\ldots,c_c) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m d_{ij}^2 \]  

(3)

Where, \( U_{ij} \) is between 0 and 1; c_i is the centroids of cluster i; d_{ij} is the Euclidean Distance between i-th. Centroids c_i and j-th. Data point. \( m \in [1, \infty] \) is a weighting exponent. There is no prescribed manner for choosing the exponent parameter, “m”. In practice, m=2 is common choice, which is equivalent to normalizing the coefficients linearly to make their sum equal to 1. When m is close to 1, then the cluster centre closest to the point is given much larger weight than the others and the algorithm is similar to k-Means.

To reach a minimum of dissimilarity function there are two conditions. These are given in (4) and (5).

\[ c_i = \frac{1}{\sum_{j=1}^{c} \sum_{k=1}^{n} u_{ik}^{2m-1} d_{ik}^{2m-1}} \]  

(4)

\[ u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ik}}{d_{ij}} \right)^{2m-1}} \]  

(5)

This algorithm determines the following steps in Figure2.
**FCM ALGORITHM:**

**Input:** n data objects, number of clusters  
**Output:** membership value of each object in each cluster  

**Algorithm:**
1. Select the initial location for the cluster centres  
2. Generate a new partition of the data by assigning each data point to its closest centre.  
3. Calculate the membership value of each object in each cluster.  
4. Calculate new cluster centers as the centroids of the clusters.  
5. If the cluster partition is stable then stop, otherwise go to step 2 above.

Figure 2: Fuzzy C-Means Clustering Algorithm

By iteratively updating the cluster centres and the membership grades for each data point, FCM iteratively moves the cluster centres to the “right” location within a data set. FCM does not ensure that it converges to an optimal solution, because the cluster centers are randomly initialised. Though, the performance depends on initial centroids, there are two ways as described below for a robust approach in this regard.

1) Using an algorithm to determine all of the centroids.
2) Run FCM several times each starting with different initial centroids.

More mathematical details about the objective function based clustering algorithms can be found in [11].

### 3.3. EM Clustering

Finite mixture distributions provide a flexible and mathematical-based approach to the modelling and clustering of data observed on random phenomena. We focus here on the use of the normal mixture models, which can be used to cluster continuous data and to estimate the underlying density function. These mixture models can be fitted by maximum likelihood via the EM (Expectation-maximisation) algorithm. The main assumption is that data points are generated by, first randomly picking a model j with probability $\tau_j$, $j=1: k$, and, second, by drawing a point x from a corresponding distribution. The area around the mean of each (supposedly unimodal) distribution constitutes a natural cluster. So, we associate the cluster with the corresponding distribution’s parameters such as mean, variance, etc. Each data point carries not only its (observable) attributes, but also a (hidden) cluster ID (class in pattern recognition). Each point x is assumed to belong to one cluster, and we can estimate the probabilities of the assignment $Pr(C_i|x)$ to jth model. The overall likelihood of the training data is its probability to be drawn from a given mixture model

$$L(X|C) = \prod_{i=1}^{N} \sum_{j=1}^{k} \tau_j Pr(X_i|C_j)$$  \hspace{1cm} (6)

Log-likelihood $log(L(X|C))$ serves as an objective function, which gives rise to the Expectation-Maximisation (EM) method. EM is a two step iterative optimization. Step (E) estimates probabilities $Pr(x|C_j)$, which is equivalent to a soft (fuzzy) reassignment. Step (M) finds an approximation to a mixture model, given current soft assignments. This boils down to finding mixture model parameters that maximise log-likelihood. The process continues until log-likelihood converges is achieved.

Because the mixture model has clear probabilistic foundation, the determination of the most suitable number of clusters k becomes a more tractable task. From a data mining perspective, excessive parameter set causes over fitting, while from a probabilistic perspective, number of parameters can be addressed within the Bayesian framework. More details can be found in [12].
A Comparative Study of Clustering Algorithms for Building a Network Intrusion Detection Model

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t densities is that it implies clustering is invariant under affine transformations of the data (that is, under operations relating to changes in location, scale, and rotation of the data). Thus, the clustering process does not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space.

As with k-means, the EM algorithm is only guaranteed to converge to a local maximum, not the global one and so the procedure should be repeated a number of times with different initial guesses for the parameter values. In this case, however, the log-likelihood figure can be used to directly compare the final configurations obtained and so the user just has to choose the largest of the local maxima.

The standard EM algorithm generates a series of parameter estimates, where represents the reaching of the convergence criterion, through the following steps, as shown in figure 3.

The major disadvantages for EM algorithm are the sensitivity to the selection of initial parameters, the effect of a singular covariance matrix, the possibility of convergence to a local optimum, and the slow convergence rate [13]. Variants of EM for addressing these problems are discussed in [13] and [14]. A valuable theoretical note is the relation between the EM algorithm and the K-means algorithm. Celeux and Govaert proved that classification EM (CEM) algorithm under a spherical Gaussian mixture is equivalent to the K-means algorithm [15, 16].

4. CLUSTERING EVALUATION SCHEMES

The result of the cluster analysis can be evaluated with the help of any of the following, in order to measure their effectiveness in building an network intrusion detection model.

4.1. Objective Function/Log Likelihood Criteria

Usually for clustering, there are two kinds of measures of cluster “goodness” or quality [17]. One type of measure allows us to compare different sets of clusters without reference to external knowledge and is called an internal quality measure. This type of measure uses the similarity of documents in a cluster. The second type of measure evaluates how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an external quality measure. One external measure is entropy, which provides a measure of “goodness” for un-nested clusters or for the clusters at one level of a hierarchical clustering. Another external measure is the F-measure, which is more oriented towards measuring the effectiveness of a hierarchical clustering. For EM algorithm, we may use “log likelihood” to measure the “overall similarity”, since in EM, in each iteration, we optimize the log likelihood of expected parameters.

Figure 3 : Expectation-Maximization (EM) Algorithm
The algorithm terminates when a formula that measures cluster quality no longer shows significant increases. One measure of cluster quality is the likelihood that the data came from the dataset determined by the clustering. The likelihood computation is simply the multiplication of the sum of the probabilities for each of the instances. With two clusters A and B containing instances $x_1, x_2, \ldots, x_n$, where $P_A = P_B = 0.5$, the computation is:

$$[0.5P(x_1|A) + 0.5P(x_1|B)] + [0.5P(x_2|A) + 0.5P(x_2|B)] + \ldots + [0.5P(x_n|A) + 0.5P(x_n|B)]$$  \hspace{1cm} (7)

The log likelihood score is the logarithm of the likelihood measure defined above in equation (7). There is no upper bound on this value; however, larger scores represent clustering of higher quality.

In case of FCM, the minimum value of the objective function is equal to twice the logarithm of the likelihood. If the likelihood is smaller than 1, the log will be negative. The likelihood, in simple normal problems is a sum of squares. If that sum is >1, then -2 log likelihood will be negative. The absolute value of the objective function is meaningless, as a likelihood is only defined up to an arbitrary proportionality constraints. Only differences between objective functions of nested models are meaningful.

4.2. Using Number of Clusters

EM allows us to choose the number of clusters to be formed. As an alternative, it can also be possible to instruct EM to determine a best number of clusters. The algorithm will converge to an optimal clustering; however, the optimization may not be global.

4.3. Choice of Seeds

The clustering results can vary based on random seed selection. Some seeds can result in poor convergence rate or convergence to sub-optimal clustering. So, the choice of seed is of paramount importance in good clustering evaluation in order to build an efficient intrusion detection model.

5. EXPERIMENTAL SETUP AND RESULTS

In this experiment, we have used a standard dataset, the raw data used by the KDD Cup 1999 intrusion detection contest [2]. However, in our experiment; we have used 10% KDD Cup’99 datasets. This database includes a variety of intrusions simulated in a military network environment that is common benchmark for evaluation of intrusion detection techniques. This data set consists of 65525 data instances, with 21 training attack types, each of which is a vector of extracted feature values from a connection record obtained from the raw network data gathered during the simulated intrusion and is labelled as either normal or a certain attack type. The distribution of attacks in the KDD Cup’99 dataset is highly unbalanced. Some attacks are represented with only a few examples, e.g. the phf and ftp_write attacks, whereas the Smurf and Neptune attacks cover many records. In general, the distribution of attacks is dominated by probes and DoS attacks.

We carried out the experiments on 2.8GHz Pentium IV processor, 512 MB RAM running Windows XP system. Weka tool [18] was used for performing the K-Means and EM clustering experimentation, where as Fuzzy Logic Toolbox [19] of MATLAB 7.0 was used for fuzzy c-Means clustering. The K-Means algorithm finds k clusters by choosing k data points at random as initial Cluster centers. Each data point is then assigned to the cluster with center that is closest to that point. Each cluster center is then replaced by the mean of all the data points that have been assigned to that cluster. This process is iterated until no data point is reassigned to a different
cluster. The simulation result of K-Means algorithm is shown in figure 4, in terms of Receiver Operating characteristics, which is a measure between Detection rate and False Alarm Rate. This is a good indicator of performance, since it measures what percentage of intrusions the system is able to detect and how many incorrect classifications are made in the process.

In practice, the number of classes is not always known beforehand. There is no general theoretical solution to find the optimal number of clusters for any given dataset. We choose k=5 for the experimentation of the FCM. The simulation results after using FCM are shown in figures 5, and 6. In figure 7, the shape of the membership function for selected values of the fuzzification factor (m=2) and cluster number=5 is shown. It can also be seen from these figures that, we are able to group the data by using the objective functions based fuzzy c-means clustering approach. Finally, the relationships of the objective function with the number of iterations are obtained in figure 8.

In case of EM Clustering, the aim is to select the number of clusters automatically by maximizing the logarithm of the likelihood of future data, estimated using cross-validation. Beginning with one cluster, it continues to add clusters until the estimated log-likelihood decreases. The results are illustrated in figures 9, 10, 11, and 12 based on various criteria as illustrated in following section.

6. DISCUSSION
The K-means clustering algorithms are the simplest methods of clustering data. The K-means algorithm uses a set of unlabeled feature vectors and classifies them into k classes, where k is given by the user. From the set of feature vectors k of them are randomly selected as initial seeds. The feature vectors are assigned to the closest seeds depending on its distance from it. The mean of features belonging to a class is taken as the new center. The features are reassigned; this process is repeated until convergence. The effectiveness of the K-Means clustering is shown in figure 4.
The fuzzy based clustering methods had shown tremendous achievements in areas of image processing and pattern recognition. The fuzzy c-means is a good choice for circular and spherical clusters, but if the orientation of natural clusters is not spherical, then the algorithm leads to among almost wrong clusters. Another drawback of the algorithm is that it imposes equal size clusters on the data set which is again a deviation from the natural clusters. The performance of any fuzzy based clustering method is the best when the number of clusters is known Apriori. But most of the time, it is not the case and so researchers have devised a number of methods known as cluster validation indices to evaluate the clusters formed [20, 21]. The simulation results of the FCM are shown in figures 5, 6, 7, and 8.

The reason we are using EM is to fit the data better, so that clusters are compact and far from other clusters, since we initially estimate the parameters and iterate to find the maximum likelihood for those parameters. EM uses the Maximum likelihood, in which it assumes that the parameters are fixed; the best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed. In most cases the observed data could be the samples that are used for training. Based on the number of Seeds and number of Clusters used, the effectiveness of the EM clustering is shown in figures 9 and 10 respectively. Comparison of K-Means and EM clustering is made with the help of ROC is shown in figure11, which shows that EM is a better choice in comparison to K-Means clustering algorithm. Finally, the comparison of K-Means, FCM, and EM algorithms are made with the help of their objective function in figure12. This result shows that, the EM algorithm is more suitable in comparison to all others in order to build a efficient intrusion detection model.
A Comparative Study of Clustering Algorithms for Building a Network Intrusion Detection Model

Figure 11: Comparison of K-Means and EM Clustering Based on ROC

Figure 12: Objective Function Based Comparison of Clustering Algorithms

7. Related Work

In [22], a speed up technique for image data was proposed. In this method, FCM convergence is obtained by using a data reduction method. Data reduction is done by quantization and speed-up by aggregating similar examples, which were then represented by a single weighted exemplar. The objective function of the FCM algorithm was modified to take into account the weights of the exemplars. However, the presence of similar examples might not be common in all data sets. They showed that it performs well on image data sets. However, the above algorithm does not address the issue of clustering large or very large datasets under the constraints of limited memory.

Recently in [23], a sampling based method has been proposed for extending fuzzy and probabilistic clustering to large or very large data sets. The approach is based on progressive sampling, which can handle the non-image data. However, the termination criteria for progressive sampling could be complicated as it depends upon the features of the data sets.

In [24], two methods of scaling EM to large data sets have been proposed by reducing time spent in E-step. In the first method, which is referred to as incremental EM, data is partitioned into blocks and then incrementally updating the log-likelihoods. In the second method, lazy EM, at scheduled iterations the algorithm performs partial E and M steps on a subset of data. The methods used to scale EM may not generalize to FCM as they are different algorithms with different objective functions.

8. Conclusion

The applications of fuzzy based methods in all fields of engineering and sciences have shown far reaching results and their applications in intrusion detection are also optimistic. In this paper, we have discussed the objective function based fuzzy c-means clustering in detail and their application in detecting anomaly based network intrusions. Fuzzy clustering leads to information granulation in terms of fuzzy sets or fuzzy relations. Membership grades are important indicators of the similarity of patterns or their deviance from the cluster. The advantage of using fuzzy logic is that it allows one to represent concepts where objects can fall into more than one category (or from another point of view, it allows representation of overlapping categories). The results obtained in this paper show that FCM works very efficiently in obtaining compact well separated clusters to detect network intrusions. Though we have already seen many examples of successful application of cluster analysis, there still remain many open problems due to the existence of many uncertain factors. These problems have already attracted and will continue to attract intensive efforts from broad discipline.
REFERENCES


[21] Steven Eschrich, Jingwei Ke, Lawrence o. Hall and Dmitry B. Goldgof, "Fast accurate fuzzy Clustering"
A Comparative Study of Clustering Algorithms for Building a Network Intrusion Detection Model


Author’s Biography

Mrutyunjaya Panda holds a Master Degree in Engineering and is presently working as an Assistant Professor in the Department of Electronics & Tele Comm. Engg., Gandhi Institute of Engineering and Technology, Gunupur, India. He has 11 years of teaching experience. Currently, he is pursuing Doctoral research in Computer Science in Berhampur University, Orissa, India. He has about 20 publications to his credit. His research interests include Data Mining, Network Security, Intrusion Detection and Soft Computing. He is also a reviewer of an International journal (IJSDIA). He is a life member of CSI (India), IETE (India), and ISTE (India), ACEEE, MIAENG (Hong Kong).

Dr. Manas Ranjan Patra holds a Ph.D. Degree in Computer Science from the Central University of Hyderabad and is presently working as a Reader in the Department of Computer Science, Berhampur University, India. He has worked in the International Institute for Software Technology, Macao as a United Nations Fellow during 2000. He has 22 years of experience in teaching and research in different areas of Computer Science. He has about 70 international and national publications to his credit. His research areas include Software Engineering, Data Mining, Intrusion Detection, Artificial Intelligence, and e-business. He has presented papers and chaired technical sessions in many conferences. He is a member of number of professional bodies. He has executed visiting assignments to many Institutions and Universities.
Semi Supervised Ensemble Clustering Algorithm (GA Based)
For High Dimensional Genomic Data

P. Krishnakumari 1  K. Vivekanandan 2

1Senior Lecturer, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore - 641044, Tamilnadu, India. Email: kkjagadeesh@yahoo.com
2Reader, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT
Clustering high-dimensional spaces are a difficult problem which is recurrent in many domains, for example in computational biology. Developing effective clustering methods for such domains are rare and also it is a challenging problem. This paper presents an efficient algorithm designed for high-dimensional gene data which combines the ideas of Linear Discriminant Analysis LDA based on PCA feature extraction along with K-Means algorithm to select the most discriminative subspace and it uses genetic algorithm for performing local optimization from the points that are globally optimal. The clustering process is thus integrated with the subspace selection process based on LDA and the data are then simultaneously clustered while the feature subspaces are selected. Then clustering instances are aggregated to generate final clusters based on agglomerative clustering. Also genetic algorithm is used to eliminate the problem of local optimality. Real datasets show that the proposed method outperforms existing methods for clustering high-dimensional genomic data in terms of accuracy and time.

Keywords: Gene expression, clustering, microarray analysis, K-Means clustering, Linear Discriminant Analysis, PCA, Genetic algorithm.

1. INTRODUCTION
Clustering in high-dimensional spaces is a difficult problem often referred as the “curse of dimensionality” for various application domains, such as information retrieval, computational biology, and image processing since the data dimension is usually very high for such applications. While various dimension reduction techniques have been proposed, there are two major types, feature transformation and feature selection [10][17]. Feature transformation methods project the original high dimensional space onto a lower dimensional space, while feature selection methods select a subset of meaningful dimensions from the original ones. The simplest approach of dimension reduction techniques includes principal component analysis (PCA) [9][12] and random projections [6]. In these methods, dimension reduction is carried out as a preprocessing step and is decoupled with the clustering process. Once the subspace dimensions are selected, they remain fixed during the clustering process. An extension of this approach is the adaptive dimension reduction approach [8][14][15] where the subspace is adaptively adjusted and integrated with the clustering process. Subspace clustering algorithms that detect clusters in axis parallel to the projections of the original data set [13] are not able to capture local data correlations and find clusters of correlated objects since the principal axes of correlated
data are arbitrarily oriented. If we restrict the subspace to be linear combinations of original features, the subspace obtained in linear discriminant analysis (LDA) is perhaps the best subspace to do data clustering, because in LDA subspace, clusters are well separated. Bioinformatics and genomic sequence analysis, in particular is one of the hottest topics in modern science. The usefulness of statistical techniques in this field cannot be underestimated. The increasing use of DNA microarrays to generate large-scale datasets of gene expression has led to several important statistical and analytical challenges. Microarray experiments are being carried out in biological and medical researches to address a wide range of problems, including clustering of gene data [4] [3].

In traditional partitional based algorithms, problems due to initialization and local optima do arise. Also they find difficult to handle high dimensional data. Hence an algorithm is proposed to handle high dimensional data and also eliminates local optimality. The paper is organized as follows: Section 2 presents the related work, Section 3 presents theoretical problem definition for LDA based K-means clustering, Section 4 explains the proposed GA based learning framework by combining LDA ,K-means clustering and agglomerative clustering , Section 5 presents the experimental results, and finally Section 6 provides the conclusion.

2. RELATED WORK

In 2001, LDA with K-Means is a very well developed theory with the growth of matrix-based approaches in machine learning [11][7][19][16][5]. LDA+K-Means algorithm reduces to the adaptive dimension reduction (ADM) algorithm [8] where only between-class scatter is optimized rather than the full LDA and adaptively modifies the subspace to fit the data distribution; here between-cluster scatter matrix is chosen explicitly. In 2004, LDA+K-Means algorithm reduces to the adaptive subspace iteration algorithm [14][15] where only the within-class scatter is optimized rather than the full LDA; here between-cluster scatter matrix is chosen implicitly. In 2006, a matrix factorization [7] is proposed such that, after one matrix factor is eliminated; the two remaining matrix factors can be viewed as the projection directions in a LDA variant and cluster indicators respectively. They are solved in an alternative fashion using LDA and a soft-clustering similar to adaptive dimension reduction. In 2007, LDA and K-Means clustering are simultaneously used as adaptive dimension reduction approach clustering [5] because they minimize the within-class scatter matrix and maximize the between-class scatter matrix and can be viewed as an unsupervised LDA. The proposed algorithm in this paper is developed based on this direction. In partitional clustering, problems due to initialization and local optima do arise. One way of approaching this challenge is to use stochastic optimization schemes such as genetic algorithms (Gas)[4].

The above algorithms adaptively modify the subspace to fit the data distribution when either the natural clusters in the data are close to spherical Gaussians or natural clusters are well separated. However there is a possibility of having suboptimal clusters as the clustered output directly depends on the selected dimensions for every run of the algorithm output. As a result there will be much deviation for the clustering results obtained and hence it does not guarantee for the good model for such data. Hence to modify the subspace adaptively to converge to the subspace where clusters are most separable and also to improve the clustering accuracy the proposed algorithm alternatively stores every obtained LDA
clustering instances in a similarity n x n matrix where n is the number of clustering instance generated. Based on the similarity matrix, an agglomerative clustering is finally applied to generate the final clustered result. The proposed algorithm combines LDA and K-Means in a simpler way and finally performs agglomerative clustering to improve the clustering accuracy for the high dimensional data. Thus the proposed algorithm alternatively combines selected dimensions and clustering and finally covers all the dimensions because for clinical applications coverage of all dimensions plays a major role. It is also integrated with genetic algorithm to provide global optimality.

3. Problem Definition

3.1 Linear Discriminant Analysis, C-classes: Derivation

Here (C-1) discriminant functions are used. The projection is from N-dimensional space onto (C-1) dimensions. The generalization of the within-class scatter matrix is

\[ S_w = \sum_{i=1}^{c} S_i \]

where \( S_i = \sum_{x_i} (x_i - \mu_i)(x_i - \mu_i)^T \) and \( \mu_i = \frac{1}{N_i} \sum_{x_i} x_i \)

The generalization for the between-class scatter matrix is

\[ S_b = \sum_{i=1}^{c} N_i (\mu - \mu_i)(\mu - \mu_i)^T \]

where \( \mu = \frac{1}{\sum_{x_i} N_i} \sum_{x_i} x_i \) and \( \mu_i = \frac{1}{N_i} \sum_{x_i} x_i \)

which can be arranged by columns into a projection matrix \( W = [w_1, w_2, ..., w_c] \) so that

\[ y_i = w_i^T x \Rightarrow y = W^T x \] (3)

It can be shown that the optimal projection matrix \( W^* \) is the one whose columns are the eigenvectors corresponding to the largest eigenvalues.

Since the projection is not scalar (it has C-1 dimensions), LDA produces the projection matrix \( W^* \) that maximizes the following

\[ J(W) = \frac{\tilde{S}_b}{\tilde{S}_w} = \frac{W^T S_b W}{W^T S_w W} \]

(4)

3.2 K-Means Based LDA

The standard K-means clustering minimizes the clustering objective function

\[ \min_H J_K, J_K = \sum_{k \in C_k} \sum_{i \in H_k} \|x_i - m_k\|^2 \]

(5)

where the matrix \( H = \{0, 1\}^{n \times k} \) is the cluster indicator: \( H_{ik} = 1 \) if \( x_i \) belongs to the \( k \)-th cluster, and \( H_{ik} = 0 \) otherwise. \( \text{Tr} M \) indicates the trace of matrix \( M \). It is well-known that \( S_t = S_w + S_b \). It is clear that the K-means clustering objective function is

\[ J_K = \text{Tr} S_t = \text{Tr} (S_t - S_b) \]

(6)

Therefore, K-Means clustering minimizes the within-class scatter matrix \( S_w \) or maximizes the between-class scatter matrix \( S_b \) since \( \text{Tr} S_t \) is a constant. On the other
hand, given class labels as specified by the indicator matrix H, the LDA directions U are determined by

\[
\text{max}_U \frac{\text{Tr} U^T S_b U}{\text{Tr} U^T S_b U} = \text{max}_U \frac{\text{Tr} \sum_{i \in C} U^T (x_i - m_i)(x_i - m_i)^T U}{\text{Tr} \sum_{i \in C} U^T (x_i - m_i)(x_i - m_i)^T U}
\]

Thus LDA has very similar properties as K-means clustering: minimizing within-class scatter \( S_w \) and/or maximizing between-class scatter \( S_b \). LDA is widely used to select the subspace (feature space) which has the maximal discriminant power. However, LDA is a supervised learning method which requires the class label for each data point beforehand. Since LDA and K-means clustering both minimize \( S_w \) and maximize \( S_b \), there should be ways to combine them into a single framework. In this paper, an algorithm is proposed to combine them into a single framework. K-Means clustering is used to generate class labels and use LDA to do subspace selection. The final results of this learning process are that the data are clustered while the feature subspaces are selected simultaneously and corresponding instances are stored and clustered using agglomerative clustering. LDA finds the most discriminative subspace in an unsupervised manner and optimizes the LDA objective function

\[
\text{max}_U \frac{\text{Tr} U^T S_b U}{\text{Tr} U^T S_b U} = \text{max}_U \frac{\text{Tr} U^T (S_b - S_w) U}{\text{Tr} U^T S_b U} = \frac{\text{Tr} U^T S_b U}{\text{Tr} U^T S_w U} - 1.
\]

Since \( \text{Tr} U^T S_b U \) is independent of \( H \), this leads to

\[
\text{max}_H \frac{\text{Tr} U^T S_b U}{\text{Tr} U^T S_b U} = \frac{\text{Tr} \sum_{i \in C} U^T (x_i - m_i)(x_i - m_i)^T U}{\text{Tr} \sum_{i \in C} U^T (x_i - m_i)(x_i - m_i)^T U}
\]

This is precisely the K-means clustering in the subspace \( Y = U^T X \). Once \( H \) is calculated within and between cluster scatter matrices can be computed. \( U \) is given by \( d \) eigenvectors associated with the \( d \) largest eigenvalues of the between-cluster scatter matrix \( S_b \).4. Proposed Algorithm

The proposed algorithm is based on LDA based adaptive dimension reduction approach that combines LDA and K-Means clustering, agglomerative clustering based on PCA feature extraction. As LDA and K-means clustering are optimizing the same objective function, i.e., they both minimize the within-class scatter matrix and maximize the between-class scatter matrix, it can be viewed as an unsupervised LDA. K-means clustering is used to generate class labels and use LDA to do subspace selection. The clustering process is thus integrated with the subspace selection process and the data are then simultaneously clustered while the feature subspaces are selected. Every clustering instance is stored in an \( n \times n \) similarity matrix. Cluster membership is used as the bridge connecting the clusters discovered in the subspace and those defined in the full space. With this connection, clusters are discovered in the low dimensional subspace to avoid the curse of dimensionality and the results are aggregated to form an \( n \times n \) similarity matrix where \( n \) is the number of instances. An agglomerative clustering is then applied to the matrix to produce final results.

4.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is the method used in statistics and machine learning to find the linear combination of features which best separate two or more
To aggregate multiple clustering results, the values of \( P_{ij}^{0} \)'s are averaged across \( n \) runs to obtain \( P_{ij} \), an estimate of the “probability that data point \( i \) and \( j \) belong to the same cluster”. This forms a similarity matrix. This is tested by performing ten runs of the algorithm on the synthetic data set and separated the aggregated \( P_{ij} \) values into two groups based on if data point \( i \) and \( j \) are from the same cluster. \( P_{ij} \) values are large when data point \( i \) and \( j \) are from the same natural cluster and small otherwise.

### 4.3 The Agglomerative Algorithm

To produce the final clusters from the aggregated similarity matrix \( P \), an agglomerative clustering is applied as follows:

**Algorithm:**

**Inputs:**
- \( P \) is a \( n \times n \) similarity matrix,
- \( k \) is a desired number of clusters.

**Output:**
- a partition of \( n \) points into \( k \) clusters.

**Procedure:** An Agglomerative clustering Algorithm

\[ l = n. \]

For \( i = 1 \) to \( n \)

Let \( c_i = \{x_i\} \) for \( i = 1, \ldots, n \)

Repeat

Find the most similar pair of clusters based on \( P \), say \( c_i \) and \( c_j \).

Merge \( c_i \) and \( c_j \) and decrement \( l \) by one

Until \( l \leq k \)

The similarity between two clusters is given as follows:

\[
\text{sim}(c_i, c_j) = \min_{i_k, j_l, k_j} P_{ij}
\]

When two points have very small similarity value (i.e., small possibility of belonging together according to \( P \)) the algorithm will not group them together.

### 4.2 Aggregating Multiple Clustering Results

The clustering results are aggregated into a matrix that measures the similarity between each pair of data points. Then an agglomerative clustering algorithm is applied to produce the final clusters. For each subspace selection and clustering process, if \( \theta \) represents a model then for each data point \( i \), the soft clustering results \( P(l | i, \theta), l = 1, \ldots, k \) are given, representing the probability that the point belongs to each cluster under the model \( \theta \).

can be defined [18] as the probability of data point \( i \) and \( j \) belonging to the same cluster under model \( \theta \) and it can be calculated as :

\[
P_{ij}^{\theta} = \sum_{l=1}^{k} P(l | i, \theta) \times P(l | j, \theta)
\]
4.4. The Proposed LDA Method

A unique feature in this approach is switching between the subspace (for clustering) and the original space (for LDA). The cluster indicator $H$ enables us to uniquely connect the the spaces. With this connection, clusters are discovered in the low dimensional subspace to avoid the curse of dimensionality and are adaptively re-adjusted for global optimality.

**Algorithm Steps:**

**Step 1:** Set dimension $d$ to be $K-1$ where $K$ is number of clusters

**Step 2:** Generate class label performing K-means clustering based on initial $U$ obtained from PCA and compute $H$

**Step 3:** Store the clustering instance

**Step 4:** From $H$, derive next LDA subspace $U$ and perform k-means clustering

**Step 5:** Repeat steps 3, 4, 5 until convergence

**Step 6:** Construct the $n \times n$ similarity matrix where ‘$n’ is the number of instances

**Step 7:** Perform Agglomerative clustering to produce final clusters

As $Y = U^T X$ is proportional to $P_{ij}$ aggregation, we can write $U^T X = K P_{ij}$. Since $K$ does not depend on class labels, this is effectively close to equation (5).

The proposed partitional based ensemble method is deemed to be the best at that point in the algorithm, but may not be the best globally when all information is considered and hence it is integrated with genetic algorithm to find optimal or near optimal solutions on complex, large spaces of possible solutions. In GAs, biologically inspired operators like crossover and mutation are applied to yield a new generation of strings [4] based on fitness. The process of selection, crossover and mutation continues for a fixed number of generations or till the termination condition is satisfied.

4.5 GA Based Ensemble Algorithm

```
Algorithm
{ Randomly generate K cluster centers from initial population;
  Perform proposed LDA method;
  While (Not termination condition) do
  { Evaluate fitness ();
    Select ();
    Apply crossover ();
    Apply mutation ();
    Perform proposed LDA ensemble clustering;
  }
  Show Clusters Based on the Best Cluster Centers
}
```

Fitness is determined based on the Euclidean measure. Here single point crossover with a crossover probability of 0.9 is chosen. Mutation rate is 0.1. Best individuals are selected based on roulette wheel selection. The computational complexity of the GA based algorithm is $O(gpknkt) + O(d^2 nt) + O(p1^2)$ for K-Means clustering combining LDA and aggregating clustering instances where $d$ is the dimension of data, $n$ is the number of data points, $t$ is the number of iterations, $k$ is the number of clusters, $p1$ is the number of instances, $g$ is number of generations, $p$ is the population.

5. DATA SETS

High density DNA microarray technology can simultaneously monitor the expression level of thousands of genes which determines different pathological states of the same tissue drawn from different patients.
The proposed algorithm is implemented in MATLAB to analyze well-known data sets. Two gene data sets with relatively high dimensions are chosen from the NCBI database. First data set contains drosophila melanogaster genes. The data consists of 10000 genes, of which only 8175 genes are identified to respond significantly. The second data set is the real AML-ALL leukemia data set. AML-ALL is a gene sample data set which consists of 38 bone marrow samples over 7129 probes from 6817 human genes.

6. EXPERIMENT RESULTS

The proposed algorithm is implemented in MATLAB. The fig1 shows the LDA projection in the first step.

The proposed GA based algorithm is implemented for the two real data sets and the results are tabulated in the tables 1, 2. Table 1 shows the performance of the proposed GA based algorithm in terms of time and accuracy for the two data sets. It is observed that the proposed algorithm takes less time at higher dimensions. It is observed from the Table 2 & 3 that the mean, standard deviation and coefficient of variance for GA based ensemble are less than GA based k-means with PCA and also the GA based ensemble is more consistent because the coefficient of variance is low.

![Figure 1: shows the LDA projection in the first step](image)

<table>
<thead>
<tr>
<th>No of Dimension</th>
<th>GA Based Ensemble (data set 1)</th>
<th>GA Based K-Means (data set 1)</th>
<th>GA Based Ensemble (data set 2)</th>
<th>GA Based K-Means (data set 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (Sec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>8.12</td>
<td>8.91</td>
<td>9.02</td>
<td>10.70</td>
</tr>
<tr>
<td>200</td>
<td>8.77</td>
<td>9.51</td>
<td>9.17</td>
<td>11.20</td>
</tr>
<tr>
<td>300</td>
<td>9.33</td>
<td>10.84</td>
<td>10.23</td>
<td>13.37</td>
</tr>
<tr>
<td>400</td>
<td>10.18</td>
<td>12.92</td>
<td>10.91</td>
<td>13.92</td>
</tr>
<tr>
<td>500</td>
<td>11.11</td>
<td>13.93</td>
<td>11.15</td>
<td>14.33</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.821</td>
<td>0.619</td>
<td>0.661</td>
<td>0.410</td>
</tr>
<tr>
<td>200</td>
<td>0.732</td>
<td>0.596</td>
<td>0.651</td>
<td>0.444</td>
</tr>
<tr>
<td>300</td>
<td>0.646</td>
<td>0.422</td>
<td>0.637</td>
<td>0.531</td>
</tr>
<tr>
<td>400</td>
<td>0.716</td>
<td>0.489</td>
<td>0.799</td>
<td>0.519</td>
</tr>
<tr>
<td>500</td>
<td>0.677</td>
<td>0.423</td>
<td>0.777</td>
<td>0.513</td>
</tr>
</tbody>
</table>
Table 2: Results Of The GA Based Ensemble Clustering In Terms Of Time For Data Set 1 & 2

<table>
<thead>
<tr>
<th>Results</th>
<th>GA based Ensemble (data set 1)</th>
<th>GA based K-Means (data set 1)</th>
<th>GA based Ensemble (data set 2)</th>
<th>GA based K-means (data set 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2281.11</td>
<td>2289.29</td>
<td>2432.10</td>
<td>2438.29</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.63</td>
<td>2.71</td>
<td>3.21</td>
<td>3.33</td>
</tr>
<tr>
<td>Co-efficient of Variance</td>
<td>0.1144</td>
<td>0.1178</td>
<td>0.1319</td>
<td>0.1368</td>
</tr>
</tbody>
</table>

\[ t \text{ value} = 6.8503 \quad p \text{ value} = 0.000045 \quad t \text{ value} = 4.232 \quad p \text{ value} = 0.001738 \]

Table 3: Results of The GA Based Ensemble Clustering In Terms Of Accuracy For Data Set 1 & 2

<table>
<thead>
<tr>
<th>Results</th>
<th>GA based Ensemble (data set 1)</th>
<th>GA based K-Means (data set 1)</th>
<th>GA based Ensemble (data set 2)</th>
<th>GA based K-means (data set 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2487.20</td>
<td>2480.12</td>
<td>2391.10</td>
<td>2381.10</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.96</td>
<td>3.19</td>
<td>3.03</td>
<td>3.12</td>
</tr>
<tr>
<td>Co-efficient of Variance</td>
<td>0.1190</td>
<td>0.1286</td>
<td>0.1267</td>
<td>0.1310</td>
</tr>
</tbody>
</table>

\[ t \text{ value} = 5.14 \quad p \text{ value} = 0.000438 \quad t \text{ value} = 7.27 \quad p \text{ value} = 0.000027 \]
The accuracy is calculated using Rand index. Table 1 show that the accuracy for the proposed method is high. From the Table 2 the two-tailed P value equals 0.000045 and 0.001738 for the two data sets. From the Table 3, the two-tailed P value equals 0.000438 and 0.000027 for the two data sets. By conventional criteria, these differences are considered to be statistically significant. It is observed by t-test analysis that the p value is less than 0.05 and hence there exists significant correlation between the methods. It is concluded that there is significant difference between proposed method and K-Means with PCA with respect to accuracy and time.

7. CONCLUSION AND FUTURE ENHANCEMENT
In this work, a new ensemble algorithm is proposed that handles all the dimensions efficiently. The proposed clustering algorithm based on GA has been implemented and tested successfully using MATLAB on windows operating systems. The results show that GA based
clustering provides global solution with significant results. The time complexity of the proposed algorithm is relatively less for high dimensional data and produces better accuracies. Since the proposed algorithm helps in reduction of dimension it considerably reduces the space. In future parallel computing techniques can be applied to increase the speed of GA.

REFERENCES


**Author’s Biography**

Dr. K. Vivekanandan received the Ph.D. degree in Computer science from Bharathiar University, India. He is currently Reader in Bharathiar University, India. He has a total teaching experience of 21 years. He has published 22 papers in international and national journals. He has produced 5 Ph D’s and 11 MPhil scholars in computer science. His research interests include data mining and knowledge discovery, and Management Information System.

P. Krishnakumari received the MPhil degree in Computer science from Bharathiar University, India. She is currently pursuing her Ph.D in computer science and has a total 13 years of teaching experience. At present she is a senior lecturer in the Department of computer science of Sri Ramakrishna college of Arts and Science for women, Coimbatore. She has presented 9 research papers in international and national conferences and published 6 papers in international and national journals. Her research interests include data mining and knowledge discovery, genetic algorithms, Bioinformatics, image compression, networking.
Combining Top Down Strategy With Bottom up Approach For Image Segmentation

Mamata S. Kalas1  P.P. Halkarnikar2

ABSTRACT
Image segmentation is the problem of partitioning an image into its constituent components. In wisely choosing a partition that highlights the role and salient features of each component, we obtain a compact representation of an image in terms of its useful parts. In proposing an integrated approach for image segmentation based on a generative clustering model combined with coarse shape information and robust parameter estimation. The sensitivity of segmentation solution to image variations is measured by image resampling. Top down information & bottom up approach is combined into a semantic likelihood map in the framework of Bayesian statistics.

Keywords: Image segmentation, clustering, generalization, resampling, Bayesian statistics.

1. INTRODUCTION
The semantic abstraction from pixels to objects in computer vision requires grouping low-level information into coherent groups or segments. This segmentation stage in image interpretation is of prime importance in low and mid level vision. Since it substantially reduces the information about objects. According to Thomas Zolar and Joachim, M. Baumann [1], the segmentation process is implemented by a parametric distribution Clustering framework (PDC). PDC is combined with coarse shape information. Data groups are represented by continuous mixture models for color and texture feature distribution [1].

For any learning algorithm, the problems of robustness toward small fluctuations in the data as well as the generalization of inferred solution to previous unseen instances of dataset from the chosen domain are highly relevant. Image segmentation as a learning problem requires inferring a robust partitioning of image patches with generalization to novel images of the same type. In PDC a mixture model approach to segmentation with top down information is used [2]. PDC is an integrated approach for image segmentation based on generative clustering model combined with top down information (shape information and robust parameter estimation). The sensitivity of segmentation solutions to image variation is measured by image resampling. Shape and similarity based grouping information is combined into a semantic likelihood map in the framework of Bayesian Statistics [1].

Image segmentation is the problem of partitioning an image into its constituent components. In wisely choosing a partition that highlights the role and salient features of each component, we obtain a compact representation of an image in terms of its useful parts. A major goal of image segmentation is to identify structures in the image that are likely corresponding to scene objects. As proposed by Rousson and N. Paragios [15], current
approaches to segmentation mainly rely on image based
criteria such as grey level of image regions as well as
smoothness and continuity of bounding contours or a
combination of these [14]. The region based approaches
recursively merge similar regions. “Divide-conquer”
approaches recursively split regions into distinct sub
regions. Contour based approaches emphasize the
properties of region boundaries, such as continuity,
curvature, smoothness and shape.

Boundary extraction is an important procedure for
segmentation and pattern identification purposes in
digital images, not only recognition and interpretation
tasks, but also for object classification. The gradient
operator is a widespread tool used for these purposes,
detecting local level variations that could correspond to
contours of interests.

Automated segmentation of images has been considered
an important intermediate processing task to extract
semantic meaning from pixels. In PDC continuous
mixture models for color and texture feature distributions
represent framework data groups where as individual
image sites are characterized by feature measurements.
PDC belongs to the clustering methods that share the
property that they grow pixels or small image patches
based on some measure of homogeneity of the associated
features or of connectors in feature space. PDC approach
is based on a generative model for the measured features.
The observations at a given site are assumed to be
generated by a particular Gaussian mixture model which
is characteristics for the cluster.

1.1 Motivation

1. Implementation of Parametric Distributional
Clustering Model

In order to characterize a clustering procedure the
modeler has to specify the objects, which are to be,
clustered, the nature of the features associated with these
objects and the criteria on which the grouping is based.
PDC segmentation method characterizes image parts by
mixture of Gaussians, which define prototypical
distributions for the measured features. Using an
Expectation Maximization (EM) algorithm performs
parametric distributional clustering.

2. Representation of Shape Knowledge

a. Segment the representative image of the object
in a sketchy way.

b. The distances of every pixel in a given image
to the now centered region depicting the object
of interest are computed for the image by
applying chamfer transform.

c. Apply the Gaussian probability function to this
single image.

3. Semantic Likelihood Maps

The goal of our approach to shape driven image
portioning is to utilize shape constraints in order to
satisfactorily segment images, which contain objects of
a certain semantic category.

- Key issue concerning successful application of
shape constraints in a segmentation procedure
is given by automatically identifying those
regions in an input image which are likely to
depict an object of the semantic category in
which one is interested.

- Key idea for this is to utilize Gaussian mixture
distributions in order to discern image regions
depicting the object of interest from those,
which merely display background clutter.

4. Bootstrap Sampling

Resampling techniques can be used to generate multiple
instances of the available data. Resampling approach
provides a viable means of finding the most pronounced

1456
Combining Top Down Strategy With Bottom up Approach For Image Segmentation

and presumably, the semantically important boundaries between image regions.

5. Combining Shape and Segmentation
The images are based on shape constrained segmentation come from the Corel image database [9] and Berkeley Segmentation Data Set.

Shape Constrained Segmentation
In shape-constrained image segmentation, shape constraints are obtained by applying chamfering technique (DT), interpreted as prior and denoted by $P_s$ and the posterior probability of the foreground semantic category denoted by $P_{wf}$. Both can be combined together to arrive get shape constrained image segmentation.

Probability assessment for the semantic categories into segmentation for the input image, each image site is assigned a label according to the maximum of posterior probability values for the foreground and background respectively. After computing label, one sweep of post processing step has been applied to the segmentation in which each site is relabeled. Another post processing step is applied in which all regions with area below the aforementioned threshold are eliminated.

2. RESEARCH WORK

Empirical Study
The experiments were conducted with real world datasets, where true natural clusters are known, to validate both accuracy and robustness of consensus via mixture model. We explored the datasets using Berkeley database.

In this contribution, we present a Clustering approach based on parametric distributions that are generated from Gaussian mixture model, called Parametric Distributional Clustering approach (PDC). PDC is presented as a novel approach to image segmentation. The segmentation technique is formulated as a generative model in the maximum likelihood framework. The specific choice of clustering algorithm, is dependant on the nature of the given image primitives which might be feature vectors, feature relations. Or feature histograms. We suggest replacing non-parametric density estimation via histograms by a continuous mixture model.

2.1 EM Algorithm
The EM algorithm estimates the parameters of a model iteratively, starting from some initial guess. Each iteration consists of an expectation step, which finds the distribution for the unobserved variables, given the known values for the observed variables & the current estimate of the parameters. Maximization step, re estimates to be those with maximum likelihood, under the assumption that the distribution found in the E step is correct. Once a model is specified with its parameters, and data have been collected, one is in a position to evaluate its goodness of fit, i.e., how well it fits the observed data. Goodness of fit is assessed by finding parameter values of a model that best fits the data- a procedure called “Parameter estimation”.

2.2 Hierarchical Chamfer Matching For Shape Alignment
Matching is a key problem in digital image analysis and edges are perhaps the most important low-level image features [5]. Thus good edge matching algorithms are important. The paper edited by Borgefors presents such an algorithm, the hierarchical chamfer-matching algorithm. The al-gorithm matches edges by minimizing a generalized distance between them. The matching is performed in a series of images depicting the same scene, but in different resolutions, i.e., in a resolution pyramid. Using this hierarchical structure reduces the computational load significantly. The algorithm is
reasonably simple to implement, and it will be shown that it is quite insensitive to noise and other disturbances.

Distance transform are applied to binary feature images, such as those resulting from edge detection. Each pixel is labeled with a number to represent its distance from the nearest feature pixel. The real Euclidean distance to pixels is too expensive to calculate and for most applications an estimate can be used. These include 1-2, 3-4 transforms and other more complicated approximations. In the predistance image, each non-edge pixel is given a value that is a measure of the distance to the nearest edge pixel. The edge pixels get the value zero. The true Euclidean distance is resource demanding (time, memory) to compute, therefore an approximation is used. The operation converting a binary image to an approximate distance image is called a distance transformation (DT). It is important that the DT used in the matching algorithm is a reasonably good approximation of the Euclidean distance; otherwise the discriminating ability of the matching measure, computed from the distance values, be-comes poor (G.Borgefors, [5]).

The DT used in the HCMA. This DT uses iterated local operations. The basic idea is that propagating local distances, i.e., distances between neighboring pixels, over the image, approximate global distances in the image. The propagation of local distances can be done either in parallel or sequentially. Sequential DT’s are known as “chamfer” distances, hence “chamfer matching.”

2.3 Bootstrap Re - sampling Strategy

The problem of over fitting is of major importance for all machine learning tasks, regardless of whether they are supervised, i.e., ground-truth label information is available, or unsupervised, i.e., one has to rely solely on the measured features. The learning procedure is supposed to infer the structural characteristics of a data set while avoiding representing statistical fluctuations and, thus, the measurement noise [1].

According to B. Efron and R.J. Tibshirani [10], alleviate the data problem; resampling techniques can be used to generate multiple instances of the available data. One of the most prominent techniques is the bootstrap method [10]. We will utilize the bootstrap framework in order to assess the stability of segmentation solutions generated by the sPDC approach with respect to variations in the input image data. Here, the direct application of resampling by drawing with replacement cannot be applied. This is due to the fact that some of the pixels will be drawn more than once, while others are not chosen at all. Therefore, the basic bootstrap sampling scheme will lead to images in which a certain fraction of pixels that are not selected (i.e., blank or black pixels). In order to fill the holes in the “synthesized” image, we propose randomly drawing a replacement from the $\Delta^{*}\Delta$ vicinity of that pixel in the original image data.

Algorithm For Bootstrap Resampling

```
Require: input image $I_{\text{in}}$ of size $n \times m$;
Vicinity size $\Delta$
Ensure: bootstrap $I_{\text{bootstrap}}$ generate set $B_i$ of (location, value) -
pairs by drawing $s$ times with replacement from $I_{\text{in}}$
populate $I_{\text{bootstrap}}$ with (location, value)
pairs from $B_i$
for each location $l$ of $I_{\text{bootstrap}}$ do
  randomly draw value $v$ from local $\Delta^{*}\Delta$
  vicinity of $l$ in $I_{\text{in}}$
end for.
```

1458
3. IMPLEMENTATION STEPS AND EXPERIMENTAL RESULTS

When we presented with standard image, it is first processed by feature extraction algorithm of PDC framework that is histograms of image sites are acquired. The features, which are subject to histograms in procedure, are the values of three-color channels of the original image together with the magnitudes of the Gabor filter bank.

Shape based image segmentation starts with feature extraction representations of images. These features are usually corners and edges. Standard edge and corner detection algorithms such as sobel filtering and canny edge detection can be applied to color or grey images to generate binary feature maps [15].

Shape constrained image segmentation is implemented by using MATLAB image processing tools and statistical tools. For Parametric distributional Clustering we use EM.m for one dimension we use a general purpose image database containing images from COREL and Berkeley Dataset. All images have size of 256x256 pixels.

During the implementation, we use a platform of Pentium 3.06 GHZ CPU with 1G RAM. Image database consists of wild animals went through image segmentation algorithm. The goal of this work is to provide an empirical basis for research on image segmentation and boundary detection. I have used this data for developing Shape constrained image segmentation [1].

Shape constrained image segmentation is implemented by using MATLAB image processing tools and statistical tools. For Parametric distributional Clustering we use EM.m for one dimension we use a general purpose image database containing images from COREL and Berkeley Dataset. All images have size of 256x256 pixels.

The input data for the PDC based approach to image segmentation are histograms of feature values taken at image sites lying on a regular grid. The features, which were subject to histogramming procedure, are the values of three color channels of the original input image together with the magnitudes of the Gabor filter bank.

The dataset is considered to be generated by a mixture of Gaussian mixture models, where the cluster probabilities p(.) denotes the mixing coefficients of the model. By virtue of the generative model, we can derive clustering objective via a maximum likelihood approach. The Expectation Maximization Algorithm addresses the problem of determining the values of the free parameters for a given dataset. The Mappings of image sites to clusters is done in E Step, whereas the parameters for the continuous mixture models are fitted in the M-Step.

The data from feature extraction is fed to an EM.m program to perform PDC. Once all images are extracted, EM will perform Parametric Distributional Clustering. PDC belongs to the category of segmentation techniques.
3.2 Shaping Model

To integrate shape knowledge into the segmentation process, the problem of adequate representation has to be addressed. Although the method of shape-constrained segmentation, which is presented here, demonstrates very generic characteristics, its application context covers the identification of a wild cat in images of its natural environment. For real-world applications, it is evident that images not only contain instances of objects of interest, but also large amounts of background pixels. This background is usually composed of clutter with few discernable shape properties; it can embody a broad variety of different distributions of elementary features. Therefore, one key issue concerning the successful application of shape constraints in a segmentation procedure is given by the ability to automatically identify those regions in an input image, which are likely to depict an object of the semantic category in which one is interested.

3.3 Gaussian Mixture Model

The most important class of finite mixture models are Gaussian mixtures. The reason for the importance and widespread use of Gaussian mixtures are incidental, but include the fact that a Gaussian has a simple and concise representation requiring only two parameters: the mean \( \mu \) and the covariance \( \Sigma \).

To set a proper number of objects per image during PDC, we compute the Gaussian distributions for the parameters mean, and covariance, & MLE (Maximum Likelihood Computation). The PDC segmentation method characterizes image parts by mixtures of Gaussians, which define prototypical distributions for the measured features. Hence, it concisely summarizes the statistical properties of image regions. The key idea used here is to utilize these mixture distributions in order to discern image regions depicting the object of interest from those, which merely display background clutter.

3.4 Prior Shape Model

As a first step in the construction of a prior shape model for standing wild cat in sideward view, image is processed by a distance transform (chamfering). In the next step Gaussian probability function is applied to the distances, transforming them into probabilities while leading to a
steep decay of values in the outer regions of image. Having averaged the shape probabilities, an additional Gaussian blurring with a stencil size of 10x10 pixels is applied.

Another method, which we have implemented, for prior shape model construction is, we will start with a single object of interest in which we capture its essential shape properties, applying the distance transform and the Gaussian model to this single image. In such a way the shape constrained segmentation approach can be utilized in content-based image retrieval system with user interaction.

Distance transforms are an important tool in computer vision, images processing and pattern recognition. A distance transform of a binary image specifies the distance from each pixel to the nearest non-zero pixel. Distance transforms play a central role in the comparison of binary images, particularly for images resulting from local feature detection techniques such as edge, arc, corner detection both the chamfers [23] matching approaches make use of distance transforms in comparing binary images. Distance transforms are also used to compute the medial axes of digital shapes [23].

3.5 Bootstrap Resampling
The problem of over fitting is of major importance for all machine learning tasks, regardless of whether they are supervised, i.e., ground-truth label information is available, or unsupervised, i.e., one has to rely solely on the measured features. The learning procedure is supposed to infer the structural characteristics of a data set while avoiding representing statistical fluctuations and, thus, the measurement noise. The data at hand are assumed to originate from a stochastic source which is characterized by a statistical distribution law. Consequently, the measured feature information is considered to result from a sampling process.

To alleviate the data problem, re sampling techniques can be used to generate multiple instances of the available data. One of the most prominent techniques is the bootstrap method [1]. We will utilize the bootstrap framework in order to assess the stability of segmentation solutions generated by the sPDC approach with respect to variations in the input image data. Here, the direct application of re sampling by drawing with replacement cannot be applied. This is due to the fact that some of the pixels will be drawn more than once, while others are not chosen at all. Therefore, the basic bootstrap sampling scheme will lead to images in which a certain fraction of pixels that are not selected (i.e., blank or black pixels). In order to fill the holes in the “synthesized” image, we propose randomly drawing a replacement from the vicinity of that pixel in the original image data). Any dataset not only contains structural information about the nature of the source, but also random fluctuations. Optimally adapting the learning algorithm to the training data thus most often results in also modeling the noise.

Figure 3.5 : Prior Shape Model
Stable edges are emphasized by averaging boundaries over the set of bootstrap samples, while edge pieces that resulted from optimization artifacts or intensity fluctuations are diminished. The gain can be attributed to resampling strategy are diminished.

**Figure 3.6 : Bootstrap Resampling**

3.6 Shape Constrained Image Segmentation

Probably assessment for the semantic categories into segmentation for the input image, each image site is assigned a label according to the maximum of posterior probability values for the foreground and background respectively. After computing label, one sweep of post processing step has been applied to the segmentation in which each site is relabeled. Another post processing step is applied in which all regions with area bellow the aforementioned threshold are eliminated.

**Figure 3.7 : Shape Constrained Image Segmentation**

4. Future Enhancements And Conclusion

- Image segmentation as a learning problem requires inferring robust partitioning of image patches with generalization to novel images of the same type.
- The bottom up approach favors smooth groupings of image patches and increases the robustness of image segmentation decisions by resampling.
- The top down information flux carries knowledge of object shapes to facilitate segmentation.
- The second enhancement of PDC introduces a priori shape information as a guiding principle for segmentation.
- A set of various aspects capture a properties of the foreground objects as well as background clutter.
- The resulting posteriori probability for occurrence of an object of a specified semantic category has been demonstrated to achieve satisfactory segmentation quality on test bed images from Corel gallery.

**References**


Combining Top Down Strategy With Bottom up Approach For Image Segmentation


Author’s Biography

Mamta S. Kalas is student of M.TECH (CST) - II and doing her dissertation work on image segmentation using parametric distributional clustering. She is doing M.TECH (CST) at Shivaji University, Kolhapur of Maharashtra. She is currently working as a lecturer at Bharti Vidyapeeth College of engineering, Kolhapur. Her areas of interest are image processing, computer architecture, system programming.
An Improved DWT Domain Statistical Image Watermarking System

R. Brindha\textsuperscript{1} \quad S.C. Sharma\textsuperscript{2}

\textbf{ABSTRACT}

In this paper, an efficient blind watermark detection scheme using DWT coefficients is proposed. The embedding scheme is multiplicative and done at second level of DWT decomposition by optimum choice of the embedding strength. The detection is based on the maximum likelihood ratio method. Neyman-Pearson criterion is used to minimize the missed detection probability subject to a fixed false alarm probability. The DWT coefficients are assumed to be modeled using the Laplacian distribution. In [10], the watermark is embedded in the vertical, horizontal and diagonal subbands in the III level. In this paper, we propose to embed the watermark only in the diagonal subband in level II, which results in better imperceptibility, robustness and capacity.

\textbf{Keywords:} Watermarking, DWT, Laplacian Distribution, Neyman-Pearson Criterion, Maximum Likelihood Detection, Decision Threshold.

1. \textbf{INTRODUCTION}

Nowadays, multimedia data is stored in the digital form which makes the processing and storage easy. But this leads to unauthorized duplication of the digital data. Digital watermarking is used to solve the above problem. It deals with techniques to embed the copyright information into a digital media by making small changes in the media content.

Watermarking can be done in either spatial domain or transform domain. Spatial domain approaches like LSB technique are not content based and are simple to implement. Transform domain approaches are more robust and can be implemented adaptively. Among the transform domain techniques DCT and DWT are commonly used.

The embedding of watermark in the cover image can be done either by additive or multiplicative rule. Usually, for additive embedding, correlation detection is used to detect the watermark. Additive methods are simple and used widely. Non-additive methods are very efficient because of their ability to achieve image dependent embedding.

The history and the basic principles of watermarking are discussed in [7], [9] and [13]. The attacks and benchmarks of performance are discussed in [14]. Cox et al in [6] compares watermarking with digital communication. A general watermarking framework and its demands are discussed in [4]. Typical watermark embedding issues and retrieval difficulties are described in [2] and [8] respectively.


\textsuperscript{1}Department of ECE, Avinashilingam University for Women, Coimbatore.

\textsuperscript{2}Vice chancellor, Tumkur University, Karnataka.
third level decomposition is employed and subbands in level 3, namely, horizontal (LH3), vertical (HL3) and diagonal (HH3) are embedded with watermarks. In this paper, we use the level 2 decomposition and embed only in diagonal subband (HH2). This improves the capacity, imperceptibility and robustness.

2. METHODOLOGY

a) Embedding Scheme [10]

Let \( X = \{x_1, x_2, \ldots, x_N\} \) and \( Y = \{y_1, y_2, \ldots, y_N\} \) be the vectors representing DWT coefficients of cover image and watermarked image in the HH2 region. A watermark \( W = \{w_1, w_2, \ldots, w_N\} \) which is chosen from a set \( M \), is embedded into \( X \) giving \( Y \). \( W \) is inserted into the \( X \) by using multiplicative rule,

\[
y_i = x_i (1 + \alpha w_i) \quad i = 1, 2, \ldots, N
\]

where \( \alpha \) is the embedding strength and \( x_i, w_i \) and \( y_i \) are the values of the random variable \( X_i, W_i \) and \( Y_i \) respectively for \( i = 1, 2, \ldots, N \). The elements of the watermarks from the set \( M \) are independent and uniformly distributed in the interval \([-1, 1]\).

b) Maximum Likelihood Detection [10]

If \( W^* = \{w_1^*, w_2^*, \ldots, w_N^*\} \) is the embedded watermark, we can write \( M = M_0 \cup M_1 \), where \( M_0 = \{W : W \neq W^*\} \) and \( M_1 = \{W^*\} \). The null watermark \( W = \{0, 0, \ldots, 0\} \), which indicates that no watermark is embedded, is already included in \( M_0 \).

Two hypothesis can be established as follows:

- \( H_0 : Y \) has \( W^* \)
- \( H_1 : Y \) does not have \( W^* \)

The statistical decision test or watermark presence detection test is interpreted as deciding if the input of the detector is the outcome of the random process with the pdf conditioned to \( H_1 \) and \( H_0 \). It compares the ratio between the pdf conditioned to \( H_0 \) and the pdf conditioned to \( H_1 \) against a threshold as given below.

If the likelihood ratio,

\[
l(y) = \frac{f_Y(y|M_1)}{f_Y(y|M_0)} > \lambda
\]

then the watermark \( W^* \) is detected

where \( f_y(y|M_j), j = 0, 1 \) are the conditional pdfs and \( \lambda \) is the decision threshold. Since \( \alpha < 1 \), from [4]

\[
f_y(y|M_j) = f_y(y/0)
\]

Assuming that the transform coefficients are statistically independent, (1) can be expressed as

\[
l(y) = \prod_{i=1}^{N} f_{Y_i}(y_i/w_i^*)
\]

where \( f_{X_i} = \prod_{i=1}^{N} f_{X_i}(y_i) \) and \( f_{Y_i}(y_i/w_i^*) \) are the conditional pdfs.

Since \( \log x \) is an increasing function of \( x \), \( \log l(y) \) will reach its maximum value when \( l(y) \) reaches its maximum. Hence, taking natural log on both sides

\[
z(y) = \sum_{i=1}^{N} \ln f_x\left(\frac{y_i}{1 + \alpha w_i^*}\right) - \ln f_x\left(y_i\right) > \lambda
\]

where \( \lambda = \ln \lambda' = \sum_{i=1}^{N} \ln (1+\alpha w_i^*) \)

is the modified decision threshold.

c) Decision Threshold [10]

The Neyman-Pearson criterion is used to find \( \lambda' \) to minimize the missed detection probability for a fixed false alarm probability, \( P_{fa} \).

Let, \( P_{fa} = 10^{-6} \).
An Improved DWT Domain Statistical Image Watermarking System

\[ P_{fa} = P(z(Y) > \lambda'/M_o) = P(z(X) > \lambda') \]

As the number of \( Z(x) \) is more than 30, central limit theorem can be applied and PDF of \( Z(x) \) can be assumed to be Gaussian.

Thus,

\[ P_{fa} = \int_{\lambda'/\sigma_{z(x)}}^{\infty} \frac{1}{\sqrt{2\pi\sigma_{z(x)}^2}} e^{-\frac{(z(x) - \mu)^2}{2\sigma_{z(x)}^2}} dz(x) \]

which gives

\[ \lambda' = \text{erfc}^{-1}(2P_{fa}) \sqrt{2\sigma_{z(x)}^2} + \mu_{z(x)} \] (9)

d) Laplacian Model [10]

Each of the DWT coefficients is modeled by the Laplacian PDF given below

\[ f_{X_i}(x_i) = \frac{b_i}{\sigma_i} \exp\left( -\frac{|x_i - \mu_i|}{\sigma_i} \right) - \infty < x_i < \infty \] (10)

with \( b_i = \sqrt{2/\sigma_i} \) where \( \sigma_i^2 \) is the variance of \( X_i \) and \( \mu_i \) is the mean of \( X_i \). Substituting (10) in (4),

\[ z(y) = \left| \sum \left[ y_i - \mu_i \right] - \left[ 1 + \alpha w_i^1 \right] \left| \left[ y_i - \mu_i \right] - \mu_i \right| + \lambda' \right| \] (11)

Mean and variance are derived to be

\[ \mu_{z(x)} = \sum_{i=1}^{N} 2 \left( |1 + \alpha w_i^1| \mu_i \right) \] (12)

and

\[ \sigma_{z(x)}^2 = \sum_{i=1}^{N} \left[ |1 + \alpha w_i^1| \right]^2 \left( 2 - \exp(-b_i |\mu_i \alpha w_i^1|) \right) \]

Substituting (12) and (13) in (9) the decision threshold \( \lambda' \) is obtained.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

Images of Lena, Cameraman and Crowd at the size of 512 x 512 are used as cover images and are shown in Fig. 1. Lena contains little detail; Cameraman contains an intermediate amount of detail and crowd contains a large amount of detail [5].

\[ \text{Lena} \quad \text{Cameraman} \quad \text{Crowd} \]

Figure 1: Cover Images

Each cover image is transformed by DWT. A Daubechies filter is used to obtain a second and a third level decomposition.

3.1 Results of level III Embedding

In [10], embedding is done in the high frequency subbands LH3, HL3 and HH3. Total number of coefficients after combining the three bands is \( N = 12,288 \). Thus, watermark can have a maximum of 12,288 elements only. Each element has been assumed to be either 1 or -1. If a coefficient belongs to the particular band, mean \( \mu_i \) and variance \( \sigma_i^2 \) are estimated from the equations,

\[ \mu_i = \frac{1}{N_B} \sum_{i=1}^{N} y_i \] (14)

\[ \sigma_i^2 = \frac{1}{(N_B - 1)} \sum_{i=1}^{N} (y_i - \mu_i)^2 \] (15)

where number of coefficients in one band, \( N_b = 4096 \) and \( y_i \) is the value of DWT coefficient in band B of the
watermarked image. The results of detection are listed in Table 1 & 2. Let $\alpha = 0.3$.

Gaussian noise has zero mean and variance 0.5. Blurring is caused by circular filter of the size 31 x 11. Rotation is up to $10^\circ$ in the counter clockwise directions. JPEG compression is done to offer 50% quality. Cropping is done to obtain an image whose size is 300 x 300. Mean filter filters the image by using adaptive wiener filter, using neighbourhoods of size 4 x 4.

Table 1 : Results of level III Embedding (without attacks)

<table>
<thead>
<tr>
<th>Cover image</th>
<th>PSNR for $\alpha=0.3$ (Peak Signal to Noise Ratio)</th>
<th>Number of successful detections for 10 trials (without attacks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>37.24</td>
<td>10</td>
</tr>
<tr>
<td>Cameraman</td>
<td>34.93</td>
<td>10</td>
</tr>
<tr>
<td>Crowd</td>
<td>30.89</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2 : Results of level III Embedding (with attacks)

<table>
<thead>
<tr>
<th>Image</th>
<th>Number of successful detections for 10 trials (with attacks)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gaussian noise</td>
</tr>
<tr>
<td>Lena</td>
<td>10</td>
</tr>
<tr>
<td>Cameraman</td>
<td>10</td>
</tr>
<tr>
<td>Crowd</td>
<td>10</td>
</tr>
</tbody>
</table>

The embedded watermark is chosen from a set of 100 randomly generated watermarks of length N. Number of trials is 10. If a value increases beyond 0.3, the robustness will improve and number of successful detections will be more. But, the PSNR value which is already less will decrease still.

3.2 Results of Proposed Level II Embedding

In our proposed method, embedding is done only at HH2 subband because its variance is the lowest.

1. Capacity

HH2 subband’s size is bigger than that of HL3, LH3 & HH3 put together. Hence, the capacity of the proposed method is 16,384, whereas, the capacity of III level embedding in HL3, LH3 & HH3 is only 12,288. Thus, capacity of the proposed method is better.

2. Imperceptibility

PSNR is a measure of imperceptibility. If the embedding strength a increases, imperceptibility will reduce and robustness will improve. Similarly, if a decreases, imperceptibility will improve and robustness will reduce. Thus, there is always a trade off between imperceptibility and robustness.

In Table 3, PSNR values are compared. It is evident that the proposed level II embedding exhibits better
imperceptibility because only high frequency information is subjected to distortion in our case.

### Table 3: Comparison of PSNR Values of Existing and Proposed Methods

<table>
<thead>
<tr>
<th>Cover image</th>
<th>α</th>
<th>PSNR value of existing level III embedding</th>
<th>PSNR value of proposed level II embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>0.5</td>
<td>32.00</td>
<td>44.97</td>
</tr>
<tr>
<td>Cameraman</td>
<td>0.5</td>
<td>30.49</td>
<td>43.64</td>
</tr>
<tr>
<td>Crowd</td>
<td>0.5</td>
<td>26.49</td>
<td>40.63</td>
</tr>
</tbody>
</table>

### 3. Robustness

Number of successful detections, when the watermark image is under attack, indicate the amount of robustness. A watermark chosen from a set of 100 watermarks is embedded and detected. Table 4 and 5 show the number of successful detections for 10 trials without and with attacks. The numbers in Table 5 indicate higher robustness.

As the imperceptibility of the proposed II level embedding is high, we have chosen to embed with $a = 0.5$ which yields better robustness.

### Table 4: Results of Level II Embedding (Without Attacks)

<table>
<thead>
<tr>
<th>Cover image</th>
<th>PSNR for $a=0.3$</th>
<th>Number of successful detections for 10 trials (without attacks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>44.97</td>
<td>10</td>
</tr>
<tr>
<td>Cameraman</td>
<td>43.64</td>
<td>10</td>
</tr>
<tr>
<td>Crowd</td>
<td>40.63</td>
<td>10</td>
</tr>
</tbody>
</table>

### Table 5: Results of Level II Embedding (With Attacks)

<table>
<thead>
<tr>
<th>Cover image</th>
<th>Number of successful detections for 10 trials (with attacks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>9 10 10 10 10 10 10 10 10</td>
</tr>
<tr>
<td>Cameraman</td>
<td>9 10 10 10 10 10 10 10 10</td>
</tr>
<tr>
<td>Crowd</td>
<td>10 10 10 10 10 10 10 10 10</td>
</tr>
</tbody>
</table>

Comparing Tables 1 & 2, with Tables 4 & 5, we can conclude that level II embedding yields better robustness.

### 4. Conclusion

A maximum likelihood detection scheme based on Laplacian modeling of coefficients of DWT transformation is implemented. The results obtained at level II, HH2 subband embedding are better than the results obtained using the existing method of embedding at level III.

### REFERENCES


**Author’s Biography**

R. Brindha is a Professor in the Department of ECE of Avinashilingam University for Women, Coimbatore. She received her B.E. (ECE) and M.E. (Applied Electronics) degrees from P.S.G. College of Technology, Coimbatore. She has published 3 papers in International Journals, 1 paper in a National Journal and 4 papers in Conferences. Her area of interest is Digital image processing. She holds the post of Head of the Department of ECE.

Dr. S.C. Sharma is presently working as Vice Chancellor of Tumkur University, Karnataka State, India. He has 10 Doctorates in different fields. He has more than 120 International Refereed Journal Publications and more than 300 presentations and publications in International Conferences. He has executed more than 100 sponsored projects and industrial consultancies.
ABSTRACT

Wireless technology eradicated the networks that use cables. Wireless networks use microwaves and other radio signals for communication. One of the basic requirements is to support seamless mobility in wireless networks. Seamless means smooth transition, such that user does not perceive any delay or interruption of service. The main objective is to design a handoff decision criterion suitable to the situation. The purpose of the handoff is to keep continuous or seamless service to mobile users through different cells or network coverage. It is consequently of special importance for wireless networks. This work uses soft and hard handoff for implementation.

Keywords: WLAN, UMTS, Handoff, MIP, MSCTP

1. INTRODUCTION

Wireless technology uses electromagnetic waves to communicate information from one point to another. Although wireless technologies have been used in specific applications for decades, wireless networks have recently become much more widespread due to better technology and lower prices. Wireless networking offers various advantages over wired connections, including mobility, connectivity, adaptability, and ease of use in locations that prohibit wiring. The major problems of wireless communication are higher error rates, lower bandwidth and more frequent spurious disconnection. As a result of these factors, communication latency rises due to higher retransmission, retransmission time-out delay increases, and more error-control protocol processing is required. Communication through radio waves presents more problem than wired communication because of potential interference from the environment [esi-topic.com]. The main challenges during integration of the two wireless services are, inter working architectures, billing issues, mobility and roaming, security/authentication, Session Persistence. Depending on how much inter-dependence is required between two wireless networks, there are two different ways of integrating the two wireless technologies called loosely coupling and tightly coupling. With seamless roaming the end user will not be aware of any change in the network they are using and will not be required to interact with the system to enable handoff [Ronan Morrissey et al 2002].

2. RELATED WORK

No existing wireless networks technology can provide high bandwidth, low latency, low power consumption and wide area data service to a large number of mobile users simultaneously. Mobile node must consider three factors when performing vertical handoff to support network applications. They are mobility, load balancing and user preference. Many micro mobility design and lower layer supported protocols have been proposed, but still there is a room for further improvements [Cheng Wei Lee et al 2005]. Lima and Tejinder describes vertical...
handoff using SCTP by having client server model (i.e. between the machine and the fixed server). They use only loosely coupled architecture. The performance metrics they considered is only handoff delay and through put. (Mohammed Jaseemuddin 2003) has proposed architecture of integrating UMTS & 802.11 WLAN which allows a mobile node to maintain data (packet switched) connection through WLAN and voice (circuit switched) connection through UMTS in parallel. (Shiao –LI Tsao et al 2002) describes three possible UMTS – WLAN inter networking strategies. They got very poor performance (200ms latency and packet loss) while the user is more than 2000. (Shaojian et al 2004) had worked on TRASH, a new transport layer seamless handoff for mobile network and they considered various aspects such as handoff signaling, location management, data transfer and security considerations using multi homing feature of SCTP.

2.1 Need for the Proposed Work

The concept of receiving Internet connectivity through the mobile devices has only recently gained public familiarity. The mobile user is demanding more from the wireless industry with respect to real time application such as voice and video and non real time applications such as data services. It is designed to supplement the concept of using WLAN network “Hotspots” to extend and enhance UMTS networks in high traffic areas. Both technologies reveal that they are complementary to each other in their characteristics.

Integrating these two different technologies by handoff can lead to significant benefits to service providers and end users [Rohan Morrissey et al 2002.]. The parameters that influence the WLAN hardware’s behaviors are indicators of the signal quality. These quality parameters are usually provided by the driver of the WLAN hardware and can be gathered easily. This leads to the main idea to continuously monitor the signal quality and upon exceeding or falling below certain thresholds, alert concerned applications. With these abilities, a mobile node can control the handoff process and perform fast handoffs. The primary goal is to design the decision criterion for vertical handoff between UMTS and WLAN coverage.

2.2 Methodology Adopted

As shown in the fig [1] WLAN in the hotspot is a closed room, where one or two entrance is available for entry and exit. So a mobile user could move out and move in only via that entrance. WLAN is placed arbitrarily anywhere inside the UMTS networks There are two possibilities of occurring handoff here. They are:

i) The mobile user is moving towards the entrance of the hotspot from the WLAN. Instantly the signal strength of the mobile node is getting decreased. So, an inter technology handoff is initiated by the mobile node between WLAN and UMTS networks.

ii) A mobile user from UMTS is moving towards the hotspot entrance. Even if the signal strength is strong enough, in order to utilize the bandwidth of the network, it is necessary to have a handoff from UMTS to WLAN.

To solve this problem a handoff decision criterion is designed for the mobile users.

2.3 The criteria for WLAN to UMTS handoff are:

1. The mobile user is in WLAN network.
2. RSS is measured.
3. If RSS>=Th, then we wait for T seconds
else vertical handoff from WLAN to UMTS network is initiated.
4. We go to step 2.

The above criteria can be made more adaptive by anticipating the RSS in the next sampling time and making the next sample earlier.

1. RSS (n) is measured.

2. Is RSS (n) < Th then vertical handoff from WLAN to UMTS network is initiated.//Where, RSS(n) is the RSS measured at the n th sampling

else if (2 * RSS(n) - RSS(n-1) < Th) then we wait for T_ad = T (RSS (n-Th)/RSS (n)-RSS (n-1)). //Where, ‘T_ad’ is the ad hoc sampling period for the next observation of RSS

else we wait for T seconds.

3. go to step 1

Procedure for estimating sampling time T

i) Estimated maximum velocity is V_max .

ii) Estimated minimum distance is d_min.

iii) Proposed sampling time T < V_min/V_max.

iv) Frequency of sampling is 1/T.

2.4 The criteria for UMTS to WLAN handoff are:

1. Mobile user is in UMTS network.

2. If the MN receives the beacon signal from an AP, then we set the flag as 1, otherwise we set the flag as 0.

3. The flag is checked.

4. If the flag=1, inter technology handoff from UMTS to WLAN is initiated // in order to utilize the maximum bandwidth.

else continue in the UMTS network itself.

5. We wait for T1 seconds.

6. We go to step2.

We are using a dual mobile mode terminal with both WLAN and UMTS interface. The software inside the dual mobile mode can switch between WLAN and UMTS networks, depending upon the availability of networks. Here the WLAN network is within the coverage area of the UMTS networks.

When the user is in UMTS networks, and it is near to the coverage area of the WLAN (i.e. when it receives the beacon signal), the WLAN MAC layer of the dual mobile node computes the signal strength of the received signal. Even if the signal strength is greater than the threshold value, the mobile node initiates the handoff, in order to utilize the maximum bandwidth. It is called as Soft handoff.

When the mobile node moves out of the coverage area of the WLAN network, the WLAN interface detects the reduced strength of the signal and gives the information to the dual mobile node software. Immediately the mobile node initiates the handoff from WLAN to UMTS networks. It is also called as hard handoff. Now the information is accessed from the internet via UMTS networks [Siddiqui et al 2005]
Figure 1 Scenario used for the integrated environment

The specifications used in this work are given in Table 1.

<table>
<thead>
<tr>
<th>Mobility scenario</th>
<th>Vertical intersystem handoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handoff strategies</td>
<td>Seamless handoff</td>
</tr>
<tr>
<td>Handoff initiator</td>
<td>Mobile node</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Mobile IPv4, mSCTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility parameters</td>
<td>Handoff rate, velocity</td>
</tr>
<tr>
<td>Performance measures</td>
<td>Throughput, handoff delay, Packet loss</td>
</tr>
<tr>
<td>Performance metrics</td>
<td>Received signal strength</td>
</tr>
</tbody>
</table>

We have used two protocols to implement the proposed handoff criterion. They are MIP and mSCTP.

2.5 A Network Layer Handoff Mechanism

Here the method adopted is terminal initiated handoff [Cheng et al 2005.]. WLAN and UMTS are entirely different technologies. Since this work uses infrastructure based WLAN, it is directly connected to internet via access point. Based on the network layer mobility, MIP (Mobile IP) became a standard protocol for the mobile users. In view of its registration and authentication procedures, this protocol suffers significant handoff delay and packet loss. Our work introduces some enhancement procedures to increase the performance of MIP. In order to reduce the handoff delay and packet loss, the following enhancements are made:

i) A buffer of capacity 20 is introduced in CN (Correspondent node) and in the agents.

ii) When the rate of MN (Mobile Node) movement increases or if there is a ping-pong effect between WLAN and UMTS networks, the registration of MN with a home agent increases. This will causes a high handoff delay. In order to reduce this, new agent was defined [Ian et al 2004, Eva et al 2004, Hesham Soliman et al 2004.].

The steps involved will be agent discovery, authentication, registration and handoff. We have used the TCP segments for data transfer. The delay, packet loss, end to end throughput are calculated using the following formulas.

\[
\text{Handoff delay(MIP)} = \text{Agent Discovery Period} + \text{Registration period.}
\]

\[
\text{End to end Throughput(MIP)} = \frac{\text{(Total Traffic offered-Total data loss)}}{\text{Total transmission duration}}
\]

\[
\text{Packet loss (MIP)} = \frac{\text{Sum of data loss incurred by each handoff}}{\text{Packet size}}
\]

2.6 Transport Layer Seamless Handoff Mechanism

mSCTP is a transport layer handoff solution with Dynamic Address Reconfiguration (DAR) extension. It uses the multi homing features of SCTP (Stream Control Transmission Protocol), whereas two end hosts hold two transport layer connection identities at the same time.

We assume that WLAN and UMTS networks are overlapped. Initially, the mobile user is accessing the internet via the access point from the correspondent node using its own IP address called primary address IP1. As he/she is moving away from the boundary of the network, the received signal strength (RSS) gets reduced. It is
notices the mobile node using the link layer signaling or by the signal given by the correspondent node by using the previous data loss pattern of the old path. So the mobile node will get a new IP address by requesting the server or by using the DHCP protocol. This IP address is also called as secondary address IP2 and is sent to the CN using the Asconf Add IP address parameter. To avoid signaling overhead, Asconf Add IP address is sent along with set primary address. In turn, the CN will reply with Asconf Ack message. Immediately the CN will update its local routing table. Once it is updated, all packets will be routed only to the secondary IP address, because the primary IP address is not reachable [Li Ma et al. 2004, Shaojian et al. 2004].

Once the mobile node moves towards the boundary of the WLAN network, as it is mandatory to make use of the bandwidth of the local network, the mobile node will give a Asconf deactivate message to the CN instead of giving the Asconf delete IP address as the signaling may overburden the network. Now the packets will be routed only to primary IP address (IP1). The buffer of capacity 20 sizes is introduced in the correspondent node to reduce the packet loss during the handoff. So, even if the packet loss occurs, the packets will be retransmitted to the exact path. A central database is also introduced here to store all the addresses. We have used the SCTP segments for data transfer. The delay, packet loss, and end to end throughput are calculated using the following formulas.

\[
\text{Handoff delay (mSCTP)} = \text{Router Discovery Period} + \text{DAR period (Using dual homing)}
\]

\[
\text{DAR period} = (\text{Add IP} + \text{Set-primary} + \text{Asconf Ack period}) + \text{Deactivate IP period}
\]

End to end Throughput (mSCTP) =

\[
\frac{(\text{Total Traffic offered} - \text{Total data loss} + \text{DAR chunk overhead})}{\text{Total transmission duration}}
\]

Packet loss (mSCTP) = Sum of data loss incurred by each handoff/Packet size

3. Simulation Results and Discussions

To test the performance of the integrated network architecture, the work is implemented in NS2 2.26 version. The simulated results are shown in Figure 2 through 5. The handoff rate is the number of handoffs during the simulation in unit time.

![Figure 2: Packet Losses in UMTS-WLAN Network](image)

The Figure 2 shows the number of packets loss over the handoff rate. Here, the total packet loss of MIP increases proportional to handoff rate, while that of mSCTP does not increase as that of MIP.

![Figure 3: Handoff Delay In UMTS-WLAN802.11b](image)
Figure 3 shows that, due to the registration procedure of MIP, there is an increase in delay over the handoff rate, while that of mSCTP keeps the handoff delay as very low compared to MIP. Figure 4 shows clearly that END-to-END throughput of MIP decreases as the handoff rate increases. The mSCTP maintains a constant value irrespective of the handoff rate.

![Figure 4: End To End Throughput UMTS-WLAN 802.11b](image)

The packet loss can be totally eliminated, if the buffer size is very large. On the other hand, the number of packets loss decreases as buffer size increases, which is shown in the Figure 5.

![Figure 5: Packet Losses Vs Buffer Size](image)

4. CONCLUSION

Both WLAN and UMTS can be viewed as competing and or complementary technologies. There fore an integration of WLANs with UMTS will bring advantages to the users as well as the service providers. We have considered the problem of seamless handoff of WLAN and UMTS networks. A novel scheme is proposed to improve the handoff performance. The proposed handoff scheme uses a MIP and mSCTP protocols for implementation. The analysis of performance of the proposed scheme was presented. Although currently this scheme is verified only with IEEE802.11b and UMTS networks, it does not depend on any specific protocols details. Our proposed scheme reduces the signaling overhead on the internet and minimizes packet loss for the mobile node during handoff.

REFERENCES


Health Condition Observation with High Level Efficiency in an Advanced Sensor Network

Nagarajan M1 Karthikeyan S2

1Lecturer, CMS College of Science & Commerce, Coimbatore, India. Email: mnaagarajan@gmail.com
2Assistant Professor, Caledonian College of Engineering, Sultanate of Oman. Email: skaarthi@gmail.com

ABSTRACT

This research paper mainly focuses on system architecture for smart healthcare using an advanced Wireless Sensor Network (WSN) and improving the efficiency of sensor to get the data immediately even there is patient in critical condition. It specifically targets assisted-living residents and others who may benefit from continuous, remote health monitoring. It focuses the advantages, objectives, and status of the design and multi-channel cluster clustering algorithm for increasing efficiency. An experimental living space has been constructed at the Lab for evaluation. Early results suggest a strong potential for WSNs to open new research perspectives for low-cost, ad hoc deployment of multimodal sensors for an improved quality of medical care and data retrieval time is very low.

Keywords: Wireless Network, Sensor Network, Remote Monitoring.

1. INTRODUCTION

Now a day’s people those who are suffering from diseases of the elderly will increase. In-home and nursing-home pervasive networks may assist residents and their caregivers by providing continuous medical monitoring, memory enhancement, control of home appliances, medical data access, and emergency communication.

For example, some of them are devoted to continuous medical monitoring for degenerative diseases like Alzheimer’s, Parkinson’s or similar cognitive disorders [6]. Other projects such as “CodeBlue” at Harvard extend WSNs for medical applications in disasters [7]. This network architecture for smart healthcare that will open up new opportunities for continuous monitoring of assisted and independent-living residents [9,10]. While preserving resident comfort and privacy, the network manages a continuous medical history. Unobtrusive area and environmental sensors combine with wearable interactive devices to evaluate the health of spaces and the people who inhabit them. Authorized care providers may monitor residents’ health and life habits and watch for chronic pathologies. Multiple patients and their resident family members as well as visitors are differentiated for sensing tasks and access privileges.

High costs of installation and retrofit are avoided by using ad hoc, self-managing networks. The proposed wireless system will extend healthcare from the traditional clinical hospital setting to nursing and retirement homes, enabling telecare without the prohibitive costs of retrofitting existing structures.

The architecture is multi-tiered, with heterogeneous devices ranging from lightweight sensors, to mobile components, and more powerful stationary devices. Figure 1 shows a MicaZ device from Crossbow with an environmental sensor board mounted on it. The advantages of a WSN are numerous for smart healthcare, as it provides the following important properties:

1478
a) **Portability**: Small devices collect data and communicate wirelessly, operating with minimal patient input. They may be carried on the body or deeply embedded in the environment. Unobtrusiveness helps with patient acceptance and minimizes confounding measurement effects. Since monitoring is done in the living space, the patient travels less often; this is safer and more convenient.

b) **Comfort to deploy and scalable**: Devices can be deployed in potentially large quantities with dramatically less complexity and cost compared to wired networks. Devices are placed in the living space and turned on, self-organizing and calibrating automatically.

c) **Real-time and always-on**: Physiological and environmental data can be monitored continuously, allowing real-time response by emergency or healthcare workers. The data collected form a health resource, and are valuable for filling in gaps in the traditional patient history. Even though the network as a whole is always-on, individual sensors still must conserve energy through smart power management and on-demand activation.

d) **Reconfiguration and self-organization**: Since there is no fixed installation, adding and removing sensors instantly reconfigures the network. Doctors may re-target the mission of the network as medical needs change. Sensors self-organize to form routing paths, collaborate on data processing, and establish hierarchies.

---

2. **HIGH LEVEL SYSTEM ARCHITECTURE**

A. **System Overview**

The medical sensor network system integrates heterogeneous devices, some wearable on the patient and some placed inside the living space. Together they inform the healthcare provider about the health status of the resident. Data is collected, aggregated, pre-processed, stored, and acted upon using a variety of sensors and devices in the architecture (pressure sensor, RFID tags, floor sensor, environmental sensor, dust sensor, etc.). Multiple body networks may be present in a single system. Traditional healthcare provider networks may connect to the system by a gateway, or directly to its database. Some elements of the network are mobile, while others are stationary. Some can use line power, but others depend on batteries. If any fixed computing or communications infrastructure is present it can be used, but the system can be deployed into existing structures without retrofitting.

The components of the architecture are shown in Figure 2, dividing devices into any one of several parallel layers based on their roles and physical interconnect. Each tier of the architecture is described below.

---

**Figure 1**: Mica With MTS 310 Sensor Board

**Figure 2**: Multi-Tiered System Architecture, Showing Physical Connectivity
Body Network and Subsystems: This network comprises tiny portable devices equipped with a variety of sensors (such as heart-rate, heart-rhythm, temperature, oximeter, accelerometer), and performs biophysical monitoring, patient identification, location detection, and other desired tasks. These devices are small enough to be worn comfortably for a long time. Their energy consumption should also be optimized so that the battery is not required to be changed regularly. They may use “kinetic” recharging. Actuators notify the wearer of important messages from an external entity. For example, an actuator can remind an early Alzheimer patient to check the oven because sensors detect an abnormally high temperature. Or, a tone may indicate that it is time to take medication. The sensors and actuators in the body network are able to communicate among themselves. A node in the body network is designated as the gateway to the emplaced sensor network. Due to size and energy constraints, nodes in this network have little processing and storage capabilities. More details about the particular body networks we have developed are available [10].

Emplaced Sensor Network: This network includes sensor devices deployed in the environment (rooms, hallways, furniture) to support sensing and monitoring, including: temperature, humidity, motion, acoustic, camera, etc. It also provides a spatial context for data association and analysis. All devices are connected to a more resourceful backbone. Sensors communicate wirelessly using multi-hop routing and may use either wired or battery power. Nodes in this network may vary in their capabilities, but generally do not perform extensive calculation or store much data. The sensor network interfaces to multiple body networks, seamlessly managing handoff of reported data and maintaining patient presence information.

Backbone: A backbone network connects traditional systems, such as PDAs, PCs, and databases, to the emplaced sensor network. It also connects discontinuous sensor nodes by a high-speed relay for efficient routing. The backbone may communicate wirelessly or may overlay onto an existing wired infrastructure. Nodes possess significant storage and computation capability, for query processing and location services. Yet, their number is minimized to reduce cost.

Back-end Databases: One or more nodes connected to the backbone are dedicated databases for long-term archiving and data mining. If unavailable, nodes on the backbone may serve as in-network databases.

Human Interfaces: Patients and caregivers interface with the network using PDAs, PCs, or wearable devices. These are used for data management, querying, object location, memory aids, and configuration, depending on who is accessing the system and for what purpose. Limited interactions are supported with the on-body sensors and control aids. These may provide memory aids, alerts, and an emergency communication channel. PDAs and PCs provide richer interfaces to real-time and historical data. Caregivers use these to specify medical sensing tasks and to view important data.

3. Improving Efficiency
To improve efficiency Multi Channel Cluster Algorithm can use, a typical problem that arises is the hidden node problem of neighboring nodes, which occurs when two nodes maintaining connectivity to a third node, cannot hear each other. In Fig. 3, node D is in communication with node B, where D is currently transmitting. Node C wishes to communicate with node B as well. Following the CSMA-CA protocol, node C listens to the medium, but since C does not detect node’s D transmission, it declares the medium free. Consequently, C accesses the medium, causing collisions at B.
The probability of any two nodes having hidden node relationship can go above 40% [5], representing a strong negative impact on the network throughput and packet delay, which for some applications might be critical. In IEEE.802.11 standard, the access technique has been ameliorated, recurring to the Request To Send (RTS) and Clear To Send (CTS) mechanism that reduces hidden node problem. Nevertheless, this mechanism was not introduced in the 802.15.4 standard due simplicity of the protocol. Also, the exchange of these messages results in a strong overhead in the communication, not improving significantly the throughput nor resolving the hidden node problem. As described above, the contention collision problem also rises with this access technique. The use of recursive backoffs of transmission of a node, when a busy channel is sensed, results in dropped packets, when the maximum number of backoffs, NB, is reached. This increases the average packet loss and average packet delay. Herewith is proposed an algorithm to improve performance of 802.15.4 based star topology WSNs. It recurs to a multichannel capability coordinator device, able to operate simultaneously in different radio channels as figure 4. The algorithm implemented by the coordinator is as follows:

1. All nodes, including the coordinator, use a default common radio channel for the auto-configuration procedure to establish the network infrastructure.
2. Nodes are grouped into different clusters of nodes following a certain clustering strategy described below. All nodes within each cluster can hear each other.
3. To each cluster of nodes, a specific 802.15.4 radio channel is assigned by the coordinator node to communicate with him, cluster-head of all clusters.
4. Each node is informed of the new radio-channel it shall use to communicate to the coordinator node.
5. In case new nodes appear, or some move, this algorithm may be repeated, updating the communication configuration.

The grouping strategy into clusters is described below:

a) The coordinator sends a date polling frame to each node in turn, requesting an ACK. Nodes that do not hear the ACK from that node, store that node as hidden.
b) At the end of this process, the coordinator sends a request to each node to send the list of hidden nodes. A list of nodes with their associated table of hidden nodes is build by the coordinator, ordered according to the number of hidden nodes. The node with highest Number of Hidden Nodes (NHN) is at the beginning of the list.
c) Based on this information, the coordinator builds the groups:

i. The construction of a cluster is started moving the node with higher NHN from the list to this cluster group. The second higher NHN node is then checked if it sees the first node; if it is seen it is added to the group and removed from the list, if it is hidden, it is left in the list, and the next node is checked.
ii. This process is run through the whole ordered list. For each next proponent node is checked if it is seen by all nodes in the current cluster, being added to the cluster and removed from the list if seen, or left in the list if hidden to any.

iii. When all nodes have been checked, the process restarts at i., until no more nodes are available in the list. The list of nodes and associated NHN tables is continuously updated, being removed a node from it if assigned to a certain cluster.

As proven mathematically in [2], 5 clusters are sufficient to cover all nodes around a coordinator node. The implementation of the algorithm confirms these results. As will be demonstrated, this algorithm will remove completely the hidden node problem and reduces drastically delays due to contention collision problems. Also, an increase of capacity is enabled, since 5 radio-channels of equal throughput can operate simultaneously without creating interference. Are restrictions of this mechanism that the coordinator is able to operate simultaneously in different channels, needing to have more operation capacity, and be centrally located, to be heard by every node.

4. SIMULATION

4.1 Data Acquisition

The followings are used when implement this system:

**Motion sensor:** It is a low-cost sensor module that is capable of detecting motion and ambient light levels. The module also has a simple one-button and LED user interface for testing and diagnostics. It is interfaced to a MicaZ wireless sensor node that processes the sensor data and forwards the information through the wireless network. A set of such modules is used to track human presence in every room of the simulated smart health home.

**Body network:** A wearable WSN service with MicaZ motes embedded in a jacket, which can record human activities and location using a 2-axis accelerometer and GPS. The recorded activity data is subsequently uploaded through an access point for archiving, from which past human activities and locations can be reconstructed.

**Indoor temperature and luminosity sensor:** These sensors give the environmental conditions of the habitat and are also connected to the backbone via MicaZ.

**Bed sensor:** The bed sensor, developed by the Medical Automation Research Center (MARC), is based on an air bladder strip located on the bed, which measures the breathing rate, heart rate and agitation of a patient.

**Pulse-oximeter and EKG:** These sensors were developed by Harvard University [7]. They are wearable, connecting to MicaZ and Telos devices, and collect patient vital signs. Heart rate (HR), heartbeat events, oxygen saturation (SpO2), and electrocardiogram (EKG) are available.

4.2 Backbone Infrastructure

The current backbone is a single Stargate serving as a gateway between the motes deployed in the home environment and the nurse control station. Motes use a Zigbee-compliant (802.15.4) wireless protocol for communication. The Stargate runs Embedded Linux and possesses more power and capabilities than the motes.

4.3 Database Management and Data Mining

A MySQL database serves as a backend data store for the entire system. It is located in a PC connected to the backbone, and stores all the information coming from the infrastructure for longitudinal studies and offline analysis.
4.4 Graphical User Interfaces

Interfaces with residents, healthcare providers, and technicians have different requirements. Each must present an appropriate interface for performing the intended tasks, while conforming to the constraints imposed by form factor and usability. Currently, the system offers four different GUIs.

The first is located on the local nurse control station, and it tracks the motion of the resident using motion activations. A second GUI (see Figure 5), which can run on a PDA, permits a caregiver to request realtime environmental conditions of the living space and the vital signs of the resident. It uses a query management system distributed among the PDA, Stargate and the sensor devices. The interface graphically presents requested data for clear consumption by the user. An LCD interface board was also designed for the MicaZ for wearable applications. It presents sensor readings, reminders and queries, and can accept rudimentary input from the wearer.

A final GUI, from a direct medical application based on motion sensors, exists to study the behavioral profile of the user’s sleep/wake patterns and life habits, and to detect some pathologies in the early stages.

Simulations were run in different packet sizes. The offered traffic of the totality of the nodes does not exceed the capacity of the sensor network, and in particular of the coordinator node. Results are presented in Figs. 3 and 4 for the average throughput and average packet delay. The average throughput corresponds to the fraction of traffic correctly received by the network analyzer, a device in promiscuous mode hearing all the traffic in the network. The average packet delay is the average delay experienced by a data frame from the start of its generation by the application layer to the end of its reception by the analyzer. The effect of the variation of SO and DO values was not addressed, being a detailed study on the effect of these two values presented in [19]

5. RESULTS

The system is single hop, as the radio range covers all of the facility. A multi-hop protocol will be necessary for access of multiple floors, or if transmission power is reduced. Data communication is bi-directional between the motes and the Stargate. Time-stamping is done by the PC when motion events are received. A first experiment based on seven MicaZ motes, programmed to send motion events over the network containing the location of the user, was performed with no activity in the lab for one week. We observed no false detections in the system under these conditions. However, this experiment showed the necessity of enhancing the power management scheme to prolong the lifetime of the sensors. In another experiment, the supervision program located at the control station correctly displays the
location of a mobile resident by polling the MySQL database for motion events.

6. Future Enhancement

1. Multi-modal data association and multiple residents. Data association is a way to know “who is doing what?” in a system without biometric identification and with multiple actors present, such as an assisted-living community. It permits us to recognize the right person among others when he is responsible for a triggered event. This is indispensable for avoiding medical errors in the future and properly attributing diagnostics. Consequently, dedicated sensors and data association algorithms must be developed to increase quality of data.

2. Security and privacy. The system is monitoring and collecting patient data that is subject to privacy policies. For example, the patient may decide not to reveal the monitored data of certain sensors until it is vital to determine a diagnosis and therefore authorized by the patient at the time of a visit to a doctor. Security and privacy mechanisms must be throughout the system.

7. Conclusion

The baseline of the system is implemented. The experiment showed a robust system with some straightforward communications from front to backend of the system with improved efficiency.

Recently, a vast number of wireless applications focusing on the interaction with the environment are emerging, based on Wireless Sensor Networks (WSN). IEEE 802.15.4 standard has been adopted in WSNs to achieve the wireless communication support. Nevertheless, its SMA-CA access technique suffers of the hidden node problem, strongly reducing throughput and delay of the network. An approach was presented to improve performance in star WSNs topologies. It is based on an algorithm used by a multi-channel capabilities cluster coordinator, which creates clusters of nodes that all “hear” each other. Each cluster operates in a different frequency, eliminating completely the hidden node problem, responsible for a 45% throughput decrease in single-channel configurations. From simulation results, the multi-channel strategy presents the aptitude of increasing throughput and decreasing delay by solving the hidden node problem, alleviating contention collisions and improving the global network performance.

References


Health Condition Observation with High Level Efficiency in an Advanced Sensor Network


Author’s Biography

Mr. M. Nagarajan, has completed MCA and Master of Philosophy in Computer Science at Madurai Kamaraj University. He is working as a Lecturer in CMS College of Science & Commerce, Coimbatore and doing Ph.D program in Karpagam University. He has presented three international and two national level papers

Dr. S. Karthikeyan received the Ph.D Degree in Computer Science & Engineering from the Alagappa University, Karaikudi 2008. He is Assistant Professor, Caledonian College of Engineering, Sultanate of Oman. He has published more than 15 papers in various National and International Journals. He is the Editorial Member in Journal of Computer Science. His research interests include Network Security using Cryptography and Data Mining.
Detection of Myocardial Infarction using Heart Rate Signals and Neural Networks

V. Mahesh¹  A. Kandaswamy²

ABSTRACT
Myocardial infarction (MI) is a life threatening heart disease in human beings. Early detection and treatment can save many lives. This paper presents the development of an artificial neural network, a novel nonlinear soft computing tool for the detection of myocardial infarction using heart rate data derived from ECG signals of myocardial patients and healthy subjects. This heart rate data has been used to obtain a set of statistical, spectral and spatial parameters. A feedforward backpropagation artificial neural network has been trained to predict the presence or absence of myocardial infarction on the basis of these parameters. The accuracy, specificity and sensitivity of the neural network model in identifying myocardial infarction were 95.74%, 91.67% and 100% respectively. The results demonstrate the capability of the neural network model developed in identifying myocardial infarction with significant diagnostic accuracy. The developed system can support physicians in the diagnosis of myocardial infarction.

Keywords : Artificial Neural Networks, Backpropagation, ECG, Heart Rate Variability, Myocardial Infarction.

1. INTRODUCTION
Myocardial infarction (MI) occurs when the blood supply to a part of the heart is interrupted. This is most commonly due to occlusion (blockage) of a coronary artery following the rupture of plaque, which is an unstable collection of lipids (like cholesterol) and white blood cells in the wall of an artery. The resulting ischemia (restriction in blood supply and therefore oxygen shortage), if left untreated for a sufficient period, can cause death (infarction) of heart muscle tissue (myocardium).

As millions of patients undergo health check up each year for evaluation of symptoms of MI, the demand for quality health care has thrown challenges to clinical decision making. Detection of myocardial infarction is especially challenging because it is a disease of low incidence yet a very high price has to be paid for its misdiagnosis. Techniques developed to aid the physician in his diagnosis need to be more accurate [1]. The detection of MI in today’s medical field is based on the ECG signal and traditional symptoms which include acute pain that traverses from the heart to the entire left hand. A more reliable method is to extract the Heart Rate Variability (HRV) data from the ECG signals of various patients which describes the signal in time and frequency domain. HRV analysis is a non-invasive technique that has gained prominence in the field of cardiology for detecting heart abnormalities.

The analysis of the HRV data yields various features that prove to be a better aid in the detection of heart
diseases [2,3,4] and application of artificial neural networks for disease diagnosis has yielded remarkable results [5]. This paper presents the development of a neural network based model for the detection of MI using parameters derived from HRV analysis. Subsequent sections in this paper are organized as follows: section 2 describes the Neural Network Classifier. Data for classification is dealt with in section 3, test results and discussion are presented in section 4 and the concluding remarks in section 5.

2. NEURAL NETWORK CLASSIFIER

Artificial neural networks are algorithms that are patterned after the structure of the human brain. They contain a series of mathematical equations that are used to simulate biological processes such as learning, storing and retrieving information. Neural networks have the ability to learn mathematical relationships between a series of input (independent, predictor) variables and the corresponding output (dependent, outcome) variables. This is achieved by training the network with a training data set consisting of predictor variables and the known outcome. These networks are programmed to adjust their internal weights based on the mathematical relationships identified between the inputs and outputs. Once the network has been trained, it can be used for classification of new input data known as testing data set [6]. A typical neural network is shown in Fig. 1.

The circles in this figure represent nodes and the lines connecting different nodes represent connection weights. The network consists of a set of nodes that are arranged in three layers (input, hidden and output). The input nodes are where the values of the predictor variables are presented while the output nodes represent the predicted outputs of the network. The nodes in the hidden layer contain intermediate values that are calculated by the network. Each of the hidden and output nodes contains a function called the activation function. The hidden nodes allow the network to model complex nonlinear relationships between the input and the output. In a fully connected network, each node in the input layer is connected to each node in the hidden layer, and each node in the hidden layer is connected to each node in the output layer. The knowledge gained by the learning experience through training is stored in the form of connection weights, which are used to make decisions on test inputs [7].

3. DATA FOR CLASSIFICATION

A. Source and Content

The data used in this work has been collected from the PTB Diagnostic ECG database as published in Physiobank, a site dedicated to data for various diseases and their study. This data has ECG signals of 184 subjects, of which 136 are of subjects with MI and 48 without MI. RR intervals (the time interval between two consecutive RR peaks in an ECG) were derived from these ECG signals. This RR interval data of each patient was analyzed using HRV analysis software [8], and the following parameters were obtained.

Statistical Parameters:

- Mean RR
The details of these parameters and their significance can be found in [9,12,13]. These 12 input variables and one output variable (0 for absence and 1 for presence of MI) together is called a feature vector or an instance. The dataset used in this work therefore has 184 feature vectors corresponding to 184 subjects.

B. Data Normalization

Neural network classifiers perform well with numerical data scaled to a range between 0 and 1 [6]. The input vectors were normalized so that all the values are in the desired range (0, 1).

4. RESULTS AND DISCUSSION

Different neural network models were developed and trained with 137 instances (75% of total 184 instances in the dataset) using Levenberg-Marquardt (LM), Bayesian Regulation backpropagation (BR) and Gradient descent with adaptive learning rate (GDA) algorithms for different combinations of number of hidden layers, number of neurons in the hidden layer and activation functions. Each of the trained models was tested with 46 instances (25% of total 184 instances in the dataset). The best performances of these neural network models measured through sensitivity, specificity and accuracy [10] are presented in Table 1.

From the experimental results presented in Table 1, the following were observed:

A. Performance with Different Training Algorithms

Three different back propagation algorithms such as Levenberg-Marquardt (LM) Gradient Descent with Adaptive learning rate (GDA) and Bayesian Regulation backpropagation (BR) algorithms were used for training the neural network. The best performances of these three algorithms are shown in Table 2. Fig. 2 clearly illustrates that among the LM, GDA and BR algorithms, the LM algorithm performs better [10,11] for the network chosen and the dataset used.

<table>
<thead>
<tr>
<th>SI No</th>
<th>Training Algorithm</th>
<th>Activation Function</th>
<th>No. of hidden layers</th>
<th>No. of neurons in hidden layers</th>
<th>Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GDA</td>
<td>Purelin</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>2</td>
<td>GDA</td>
<td>Logsig</td>
<td>2</td>
<td>12</td>
<td>91.489</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>3</td>
<td>GDA</td>
<td>TanSig</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>4</td>
<td>GDA</td>
<td>TanSig</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>5</td>
<td>LM</td>
<td>Purelin</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>6</td>
<td>LM</td>
<td>Logsig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>7</td>
<td>LM</td>
<td>LinSig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>8</td>
<td>LM</td>
<td>Logsig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>9</td>
<td>LM</td>
<td>TanSig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>10</td>
<td>LM</td>
<td>TanSig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>11</td>
<td>LM</td>
<td>TanSig</td>
<td>2</td>
<td>24</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>12</td>
<td>BR</td>
<td>Purelin</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>13</td>
<td>BR</td>
<td>Logsig</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
<tr>
<td>14</td>
<td>BR</td>
<td>TanSig</td>
<td>2</td>
<td>12</td>
<td>93.67</td>
<td>91.66</td>
<td>94.117</td>
</tr>
</tbody>
</table>
Table 2: Performance of Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithm</th>
<th>Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>95.744</td>
<td>91.666</td>
<td>100.000</td>
</tr>
<tr>
<td>GDA</td>
<td>89.361</td>
<td>83.333</td>
<td>97.058</td>
</tr>
<tr>
<td>BR</td>
<td>93.617</td>
<td>91.666</td>
<td>94.117</td>
</tr>
</tbody>
</table>

Table 3: Performance of Activation Functions

<table>
<thead>
<tr>
<th>Training Algorithm</th>
<th>Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>95.744</td>
<td>91.666</td>
<td>100.000</td>
</tr>
<tr>
<td>GDA</td>
<td>89.361</td>
<td>83.333</td>
<td>97.058</td>
</tr>
<tr>
<td>BR</td>
<td>93.617</td>
<td>91.666</td>
<td>94.117</td>
</tr>
</tbody>
</table>

Figure 2: Performance Comparison with Training Algorithms

B. Performance with Different Activation Functions

For the three training algorithms LM, GDA and BR different activation functions like logsigmoid, tansigmoid and purelin were used to train the network with varying number of hidden layers and number of neurons in each hidden layer. The same activation function was used for both the hidden and output layers in each model. The best results obtained for each activation function is shown in Table 3. The bar graph in Fig. 3 compares these results. The performance with logsigmoid and tansigmoid activation functions was found to be similar in two network models. However, the architecture of the model with logsigmoid activation function is simpler with one hidden layer compared to the other with two hidden layers (Sl Nos. 7 & 11 in Table 1).

C. Inference

From the experimental results presented in Table 1, it is observed that the best performance of 95.74% accuracy, 91.67% specificity and 100% sensitivity was achieved in three network models. However, the architecture 12-24-1 (Sl No. 7 in Table 1 with 12 neurons in the input layer, 24 neurons in one hidden layer, one neuron in the output layer) trained with LM algorithm and with logsigmoid activation function for both the hidden and output layer was found to be computationally efficient in terms of number of epochs required for training. The newff function in Matlab’s neural network toolbox was used for the generation of feedforward backpropagation neural network architecture. This newff function has ‘learngdm’ as the default learning function which chooses ‘intw’ initialization function for initializing a layer’s weights and biases according to the Nguyen Widow initialization algorithm. Learning occurs according to
learngdm’s learning parameters with their default values as 0.01 for learning rate and 0.9 for momentum constant. The number of neurons in the hidden layer of the architecture was considered from twelve as twelve input parameters were used in the network. Increase in accuracy, specificity and sensitivity was noted until 24 above which it led to over-fitting. A maximum of 2000 epochs were considered during the training. The convergence of the error function mse meeting the error goal of 0.001 for the best architecture (Sl Nos. 7 in Table 1) is shown in Fig. 4.

Figure 4 : Convergence of Error Function

5. CONCLUSIONS
An artificial neural network model for the detection of Myocardial Infarction has been developed. RR interval data derived from ECG signals were used with a HRV analysis software to obtain a set of parameters. These parameters were used as input to a feedforward neural network with backpropagation training algorithm to classify patients with and without myocardial infarction. The system developed does not yield results with 100% accuracy. The accuracy depends on several factors such as the size and quality of the training set, the rigor of the training imparted and also the parameters chosen to represent the input. The results presented in Table 1, indicate that the neural network classifier is effective to the tune of about 95% accuracy.

Artificial neural network techniques are initially cumbersome. It is time-intensive to collect and preprocess data and to train the networks. Once training is completed, further tasks can be carried out with relative ease. The performance of the system can be measured using real time patient data from hospitals to validate the observations. The validated system can assist physicians in the diagnosis of Myocardial Infarction.

REFERENCES
Detection of Myocardial Infarction using Heart Rate Signals and Neural Networks


Author’s Biography

V. Mahesh is an Assistant Professor in the Department of Information Technology, PSG College of Technology, and Coimbatore, India. He has a bachelor’s degree in Electronics and Communication Engineering & master’s degree in Applied Electronics from PSG College of Technology, Coimbatore. He is currently pursuing his PhD in the area of Biomedical Engineering. He has 18 years of Industrial and 7 years of teaching experience. His fields of interest are Biomedical Engineering, Microprocessors and Microcontrollers.

Dr. A. Kandaswamy is Professor and Head of Biomedical engineering Department, PSG College of Technology, Coimbatore. He has a bachelor’s degree in Electrical Engineering from PSG college of Technology, Coimbatore, master’s degree from Madras University and PhD from Bharathiar University, Coimbatore. He has 38 years of teaching experience and has published more than 80 papers in national and international journals. His fields of interest are Signal and Image processing in medicine and Wireless communication.
Complex Hadamard Matrices and Weighing Matrices

M. K. Singh1 Sudha Singh2 and D. K. Singh3

ABSTRACT
It is shown that a weighing matrix can be obtained from any quaternary complex Hadamard matrix as well as from two suitable disjoint weighing matrices. It is also shown that Hadamard and complex Hadamard matrices can be obtained from complex weighing matrices. Weighing matrices can be regarded as ternary orthogonal code. Here rows of the matrix are taken as code words.

Keywords : Dephased Hadamard Matrix, Complex Hadamard Matrix, Butson Hadamard Matrix, Conference Matrix, Weighing Matrix, Disjoint Weighing Matrices, Complex Weighing Matrix.

1. INTRODUCTION
We begin with following definitions and basic facts:
(i) A Hadamard matrix (or an H-matrix) is an nxn matrix H with entries +1, -1 such that $H^TH = nI_n$, where $I_n$ is the nxn identity matrix. If H-matrix of order n exists, n=1, 2 or 4t, where t is a positive integer.

(ii) Some generalizations of H-matrix:
(a) A complex H-matrix is an nxn matrix $H = [H_{ij}]$, where $H_{ij}$ are complex numbers with $|H_{ij}| = 1$ for $i, j = 1, 2, \ldots , n$, satisfying $HH^* = nI_n$, where $I_n$ is the identity matrix and $H^*$ denotes the Hermitian transpose of H. A complex H-matrix is called dephased or normal if elements of its first row and column are 1.
(b) A Butson H-matrix (vide Butson[5]) is an nxn complex H-matrix with elements belonging to the set of m th roots of 1 and is denoted as $BH(m,n)$. $BH(4,n)$ containing all of $\pm 1, \pm i$ is named as a quarternary complex Hadamard matrix by Horadam[11].
(c) Weighing matrix : A weighing matrix $W = W(n,w)$ of order n and weight w is an $n \times n$ (0,1,-1)-matrix such that $WW^T = wI_n$, where $W^T$ stands for transpose of W. A conference matrix of order n is a weighing matrix $W(n,n-1)$with zeroes on the diagonal.

Example : Consider the following matrix

$$
\begin{pmatrix}
0 & + & + & + & + & + & + & + \\
+ & 0 & + & + & - & - & - & - \\
+ & + & - & + & - & - & 0 & \\
+ & + & - & + & - & - & 0 & + \\
+ & - & - & 0 & + & + & + & - \\
+ & + & - & 0 & + & + & - & + \\
+ & - & 0 & + & + & - & - & + \\
\end{pmatrix}
$$

This is a symmetric weighing matrix $W(8,7)$ but not a conference matrix.
Complex Hadamard Matrices and Weighing Matrices

(d) Complex weighing matrix

A matrix $W$ of order $n$ with entries $0, \pm 1, \pm i$ is called a complex weighing matrix if $WW^t = kI_n$, where $W^t$ stands for Hermitian conjugate of $W$ and $k$ is a positive integer.

(iii) disjoint weighing matrices

Two real or complex weighing matrices $W_1 = [n_{ij}]$ and $W_2 = [n_{ij}]$, will be called disjoint if

$$m_{ij} \neq 0 \Rightarrow n_{ij} = 0$$

$$n_{ij} \neq 0 \Rightarrow m_{ij} = 0.$$


Hadamard matrices from weighing matrices were constructed by Craigen and Kharaghani [8]. For Complex weighing matrices we refer to Berman [4] and Seberry[14,15].

Complex Hadamard matrices have applications in quantum information theory and quantum tomography. Recently weighing matrices have been found much beneficial to engineers working with satellite and digital communications. They have been found to have many similarities with perfect ternary arrays, and these arrays have been implemented in our digital communications.

The purpose of the paper is to forward simple constructions for some of these matrices so that they can be used by engineers.

2 (i) Construction of a new weighing matrix from two disjoint weighing matrices:

Theorem 1: If $W_1$ and $W_2$ are disjoint nxn weighing matrices of weights $k_1$ and $k_2$ respectively, then

$$W = \begin{bmatrix} W_1 + W_2 & W_1 - W_2 \\ W_1^t - W_2^t & -W_1^t - W_2^t \end{bmatrix}$$

is a weighing matrix $W(2n, 2k_1 + k_2)$.

Proof : We have

$$WW^t = \begin{bmatrix} (W_1 + W_2)^t & (W_1 - W_2)^t \\ (W_1^t - W_2^t)^t & (W_1^t + W_2^t)^t \end{bmatrix}$$

$$= \begin{bmatrix} W_1^tW_1 + W_2^tW_2 & (W_1^t - W_2^t)(W_1 - W_2) \\ (W_1 - W_2)^t(W_1^t - W_2^t) & W_2^tW_1 + W_1^tW_2 \end{bmatrix}$$

$$= \begin{bmatrix} k_1I_n & 0 \\ 0 & k_2I_n \end{bmatrix}.$$
\[
\begin{pmatrix}
2(k_1 I_n + k_2 I_n) & 0 \\
0 & 2(k_1 I_n + k_2 I_n)
\end{pmatrix}
\]
\[
= 2(k_1 + k_2) I_{2n}
\]

Therefore, \( W = W(2n, 2(k_1 + k_2)) \).

Corollary: If \( k_1 + k_2 = n \), then \( W(2n, 2n) \) is a Hadamard matrix of order \( 2n \).

(ii) Construction of weighing matrix from the Butson Hadamard matrix \( B(4, 2n) \)

**Theorem 2** For a Butson Hadamard matrix \( B(4, 2n) \), there exists a weighing matrix \( W(4n, 2n) \) of order \( 4n \) and weight \( 2n \).

Proof: Let \( S \) be the algebra of 2x2 matrices of the form
\[
\begin{pmatrix}
a & b \\
-b & a
\end{pmatrix}
\]
over the real field \( R \). Let \( C \) be algebra of complex numbers over \( R \). Then the mapping \( \phi : C \rightarrow S \)
which carries \( a + ib \) into the matrix is an algebra isomorphism which gives faithful representation of complex number \( a + ib \) by the matrix
\[
\begin{pmatrix}
a & b \\
-b & a
\end{pmatrix}
\]
In this representation \( 1 \) can be represented by
\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]
and \( i \) can be represented by
\[
\begin{pmatrix}
0 & 1 \\
-1 & 0
\end{pmatrix}
\]
This isomorphism can be extended to the algebra of complex matrices to the algebra of real matrices by the mapping \( f \) which carries a complex matrix \( M = [a_{jk} + ib_{jk}] \) of order \( 2n \) to a real matrix \( f(M) = [\phi(a_{jk} + ib_{jk})] \) of order \( 4n \).

i.e. \( f(M) \) can be obtained from \( M \) by replacing the entry \( a + ib \) in \( M \) by the block
\[
\begin{pmatrix}
a & b \\
-b & a
\end{pmatrix}
\]
The isomorphism takes into account not only addition, usual product and multiplication by scalars but also unary operation of conjugation. Conjugate of \( (a + ib) \) corresponds to transpose of the matrix \( \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \) under the isomorphism.

In particular, if we replace \( 1 \) by \( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \), \(-1\) by \( \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \), \( i \) by \( \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \) and \(-i\) by \( \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \) in \( B = BH(4, 2n) \), we get a block matrix \( W \) of order \( 4n \) where blocks are the above 2x2 matrices. Hence \( WW^t \) is the matrix whose diagonal blocks are \( 2n \)
\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]
and non-diagonal blocks are
\[
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix}
\]
i.e. \( WW^t = 2nI_{4n} \).

Thus \( W \) is the required weighing matrix.

**Illustration**: Consider the matrix \( BH(4, 4) = \)
\[
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & i & -1 & -i \\
1 & -1 & 1 & -1 \\
1 & -i & -1 & i
\end{pmatrix}
\]
which produces
\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]
The weighing matrix \( W(4, 8) = \)

Complex Hadamard Matrices and Weighing Matrices

Gysin and Seberry [12] made nearly half-full weighing matrix conjecture which states that a weighing matrix $W(4n, 2n-1)$ exists for every positive integer $n$. They constructed a finite number of such weighing matrices. Our technique is capable of constructing some infinite classes of weighing matrices by simple replacements, supporting half-full conjecture.

3. Complex H-Matrix From Complex Weighing Matrices

Theorem 3: If there are two disjoint weighing matrices $W_1, W_2$ with entries $0, \pm 1, \pm i$ of order $n$ and weights $k_1, k_2$ such that $k_1 + k_2 = n$, then

$$
\begin{pmatrix}
W_1 + W_2 & W_1 \cdot W_2 \\
W_1^t \cdot W_2 & -W_1^t \cdot W_2^t
\end{pmatrix}
$$

is a complex Hadamard matrix of order $2n$ if $W$ stands for Hermitian conjugate of $W$.

Proof is analogous to that of theorem 1.

4. Conclusion

It is shown that a weighing matrix can be obtained from any quaternary complex Hadamard matrix as well as from two suitable disjoint weighing matrices. Since infinitely many BH$(4, 2n)$ are known (vide Turyn[17] and Wallis[18]), it follows that infinitely many weighing matrices $W(4n, 2n)$ can be constructed by theorem 2. It is also shown that Hadamard and complex Hadamard matrices can be obtained from certain disjoint weighing and complex weighing matrices respectively.

References


Author’s Biography

Prof Mithilesh Kumar Singh is University Professor at Ranchi University Ranchi. He is actively arch in various disciplines which includes Coding theory and Network Security, Mathematical modelling, Discrete and Combinatorial Mathematics and Mathematical Biology. He has Published more than 30 papers in International and National Journals. He has also visited IISc, Bangalore and JNU, New Delhi as scientist during the period 1996-98. Presently he is Director MCA Program and Head of PG Department of Mathematics at Ranchi University, Ranchi.

Sudha Singh received her M.Tech. degree in computer science from Birla Institute of Technology,Mesra,Ranchi,India, in 2002and was placed First Class First. She obtained her Ph.D. degree in Computer Science and Engineering at the BIT Sindri,(VBU),India. She has more than 10 publications. Her research interest is in recent advances in Data communication, network security, coding theory, signal processing and combinatorics.

Dr. Dharmendra K Singh is presently working as Head, Department of Electronics and Communication & Information Technology, BIT Sindri, Dhanbad. He has more than 20 years of teaching experience. He is heading the department of Electronics and Communication & Information technology since 2002. He is instrumental in starting the curriculum on information technology. He has published more than 30 papers in journals and conferences. He has already supervised 01 thesis at doctoral level in computer Science & Engg and 05 research scholars are presently enrolled under him for their doctoral degree. The area of research he works are Coding theory, cryptography, optical Amplifiers, Photonic Crystal Fibers, e-Governance and technical education Planning.