Journal of Engineering and Applied Sciences 14 (3): 831-836, 2019
ISSN: 1816-949X

Sequence Statistical Code Based Data Compression Model Using Genetic Algorithm for Wireless Sensor Networks

[1]S. Jancy and [2]C. Jayakumar
[1] Sathyabama Institute of Science and Technology, Department of CSE, Chennai, India
[2]Department of CSE, Sri Venkateswara College of Engineering, Chennai, India

Abstract: Limited capacity and limited power efficiency are considered to be the main drawback of sensors. To improve an energy efficiency of sensor is the main focus of most of the researches. Medium access control and routing protocol are the major techniques available to improve the energy efficiency of sensors. Data compression techniques are discussed in this study as a method for improving the energy efficiency of sensors through saving processing time and saving space. Genetic algorithm for data compression is proposed in this study. The assessment of SDC (Sequence Statistical Code) and FOST (First Order Static Code) are assigned as the population. The population is found followed by evaluation, selection and generation of random number. The probability of the proposed algorithm is the random number. Once, the random number is generated, Huffman algorithm is applied. The finally achieve Huffman code in cumulative probability is the final output which is the compressed data. Processing time and computation process is lesser than the Huffman algorithm with these proposed techniques.

Key words: Data compression, wireless sensor network, Huffman algorithm, Genetic algorithm, SDC, FOST

INTRODUCTION

Wireless sensor network: A sensor network is made up of several detection stations. These are called sensor nodes. The features of these sensors are that, they are light weight, small and portable. The limitations of sensors are namely limited in power, limited consumption capabilities and limited memory. Sensor nodes do not have global identification as its presence in larger amount of overhead and large number of sensors. Sensor network consist of various types of sensors. Continuous sensing event identification, event detection, location sensing and local control actuators are some of the tasks for which sensor nodes are used. .The advantages of WSN is many fold. They are able to do operate in harsh environment and without being attended. They are used in environment which is risky and not feasible for in person human monitoring. Anastasi *et al.* (2009) the major challenge is to reduce energy consumption of nodes in order to extend the life time of network to reasonable time. The requirements introduced by the existence of actors and sensors together has to be satisfies by the realization of wireless sensor and actors network (Akyildiz and Kasimoglu, 2004).

Loss and lossless data compression are the classification of data compression. Bahadili (2008) the compression of text files, codes which can be extended, word processing file and tabulation files are the areas in which lossless data compression is employed. It is also employed used in instances when the original and decompressed files should be similar. Communication is the primary field in which data compression is majorly used. This is due to its ability to facilitate devices to store or transmit the same amount of data in fewer bits. Out of the several data compression techniques which are available only few have been standardized. With OSI model data compression is the function of presentation layer. Compression algorithm can be majorly categorized in two. Lossless data compression. Lossy data Compression. The lossless data compression technique is majorly used with text data. However, lossy compression techniques are employed with audio and video data. There is no assurance for the exact same input data to be processed and send without any loss with lossy compression techniques. However, with lossless compression techniques, the exact same data is processed and transmitted without any loss.

Genetic algorithm: Genetic algorithm is predominantly used in finding solution to complex search problems. Genetic algorithm is based on evolutionary ideas of natural selection. Which is an adapted heuristic search algorithm. The stages of Genetic algorithm are initialize the population, decide the fitness of population and apply crossover function and calculate mutation and find fitness (Blanco *et al.*, 2000). Genetic algorithms are considered to be a global search technique by applying Genetic

Corresponding Author: S. Jancy, Sathyabama Institute of Science and Technology, Department of CSE, Chennai, India

operators, modelled after Genetic processes which occur in nature, the Genetic algorithm starts with a population of randomly generated solution, chromosomes and advances towards better solutions.

Algorithm 1; The processes of Genetic algorithm are:

1. The population randomly initialized (t)
2. The fitness of population is determined (t)
3. Repeat
   a) The parent from population is selected (t)
   b) Crossover is performed on parents to create population (t+1)
   c) Mutation of population is performed (t+1)
   d) fitness of population is determined (t+1)
4. Until best individuals are determined

Genetic algorithm optimization: The process of optimization is used to find an optimal solution for the given problem. The optimization problems are initiated based on three factors (Bashir, 2015).

Algorithm 2; Genetic algorithm optimization:

a) objective function
b) A set of unknowns or variables
c) A set of constraints

Literature review: By Basu *et al*. (2013) in order to achieve robust results sensor nodes in sensor network applications collaborate and share the data which is sense. Signal processing tasks are performed on the collected data and decisions are made collaborate on the sensor field events. By Tang and Raghavendra (2004) the major disadvantages of wireless sensor network is limited energy and the toughness to recharge. By Huang *et al*. (2013) in order to reduce energy consumption various data compression researches are being proposed. By Lecuire *et al*. (2007) the self adaptive image transmission scheme is proposed. This is driven by energy efficiency considerations for it to be suitable for wireless sensor networks. By Basu *et al*. (2013) an efficient bitmask compression techniques has been proposed in order to reduce the control word size considerably while the decompression over head is kept in a range which is acceptable.

Optimization problem in various fields like pattern reorganization, travelling salesman problem, machine learning, job scheduling are some of the areas in which Genetic algorithm has been applied. It starts with an initial set of random potational solutions. Which is called the populations. In order to improve them it uses a process similar to biologically evaluations (Yokota *et al*., 1996). Separation of the matching and the scheduling representation, chromosome structure independence from the communications of system details and overlap considerations among all computations and communications which obey subtask precedence constraints (Wang *et al*., 1997). Describes an enhanced bit swapping linear feedback shift register to reduce power (Hanib *et al*., 2017). Fuzzy rough gain is being used is feature selection (Moaref and Sattari-Naeini, 2017).

MATERIALS AND METHODS

Algorithm 3; Proposed algorithm (SSG):

Steps for proposed algorithm (SSG)

1. Start the process
2. Assign sequence code for given data
   Data←SDC
3. Assign First Order Static code for the data
   Data←FOST
4. Generate the Genetic algorithm
   a) Initialization
      Population←given input data
      Chromosome[data]←[SDC, FOST]
   b) Evaluation
      F_{obj} [data] = abs (SDC+2*FOST)-26)
   c) Selection
      Fitness (data) = 1/1+$F_{obj\,(data)}$
   d) Find probability
      p [I] = fitness [i]/total
   e) Take cumulative probability C (data)
   f) Generate the random number of given cumulative probability R (data)
5. Apply Huffman algorithm
   a) Random number is assigned as a proposed algorithm code
   b) Based on random number to take probability
   c) Find cumulative probability Huffman code
   d) Cumulative probability Huffman code is converted in SDC code
   e) The output of the cumulative probability is equivalent to SDC is the compressed data

Metrology of proposed algorithm (SSG): In the proposed algorithm (SSG) used Sequential Code (SDC) and First Order Statistical Code (FOST). For Example, the given input data are “ PROCEDURE”. In this input, value first to apply the Sequential Code (SDC) then to apply the First Order Statistical Code (FOST).

Sequential Statistical Code (SDC):

A-1	B-2	C-3	D-4
E-5	F-6	G-7	H-8
I-9	J-10	K-11	L-12
M-13	N-14	O-15	P-16
Q-17	R-18	S-19	T-20
U-21	V-22	W-23	X-24
Y-25	Z-26		

First Order Statistical Table (FOST):

A 0.0651738	B 0.0124248	C 0.0217339	D 0.0349835
E 0.1041442	F 0.0197881	G 0.0158610	H 0.0492888
I 0.0558094	J 0.0009033	K 0.0050529	L 0.0331490
M 0.0202124	N 0.0564513	O 0.0596302	P 0.0137645
Q 0.0008606	R 0.0497563	S 0.0515760	T 0.0729357
U 0.0225134	V 0.0082903	W 0.0171272	X 0.0013692
Y 0.0145984	Z 0.0007836	Space 0.1918182	

SDC (Sequential Code): Generate SDC Code for all alphabets:

P-16	R-18	O-15
C-3	E-5	D-4
U-21	R-18	E-5

FOST (First Order Statistical Code):

P	0.0137645	R	0.0497563
O	0.0564513	C	0.0217339
E	0.0141442	D	0.0349835
U	0.0225134	R	0.0497563
E	0.1041442		

After applying the sequential code (SDC) and First Order Statistical Code (FOST) to apply the Genetic algorithm. In Genetic algorithm first to determine the population of giving input data.

Initialization: In addition to a proper solution structure in order for Genetic algorithm search to be efficient the initial population of solutions is necessarily to be diverse representative of search space (Kwok and Ahmad, 1997; Kuo *et al*., 2001).

Example: To define the number of chromosomes in the population is 9. The generate values are {p, r, o, c, e, d, u, r, e}:

$$\text{chromosome}[p] = [16, 0.0137645]$$
$$\text{chromosome}[r] = [18, 0.0497563]$$
$$\text{chromosome}[o] = [15, 0.0564513]$$
$$\text{chromosome}[c] = [3, 0.0217339]$$
$$\text{chromosome}[e] = [5, 0.0141442]$$
$$\text{chromosome}[d] = [4, 0.0349835]$$
$$\text{chromosome}[u] = [21, 0.0225134]$$
$$\text{chromosome}[r] = [18, 0.0497563]$$
$$\text{chromosome}[e] = [5, 0.1041442]$$

RESULTS AND DISCUSSION

Evaluation: The initialization step is given by the computation of objective function values of each and every chromosome. The formula for finding the objective function:

$$F_{(obj[data]} = \text{abs}(\text{SDC}+2*\text{FOST})-26))$$

Where:

SDC = A Sequential Code for the given input data
FOST = A First Order Statistical Code for the given input data
26 = Number of possibilities of the given data

$$F_{obj[p]} = \text{abs}(16+2*0.0137645)-26) = -9.972471$$
$$F_{obj[r]} = \text{abs}(18+2*0.0497563)-26) = -7.9004874$$
$$F_{obj[o]} = \text{abs}(15+2*0.0564513)-26) = -9.870974$$
$$F_{obj[c]} = \text{abs}(3+2*0.0217339)-26) = -22.565322$$
$$F_{obj[e]} = \text{abs}(5+2*0.0141442)-26) = -20.7917116$$
$$F_{obj[d]} = \text{abs}(4+2*0.0349835)-26) = -21.930033$$
$$F_{obj[u]} = \text{abs}(21+2*0.0225134)-26) = -4.9549732$$
$$F_{obj[r]} = \text{abs}(18+2*0.0497563)-26) = -7.9004874$$
$$F_{obj[e]} = \text{abs}(5+2*0.1041442)-26) = -20.7917116$$

Selection: The next generation probability is based on the high probability of fittest chromosomes. We must compute the fitness of each chromosome for to compute fitness probability. The value of F_{obj} is added by 1 in order to avoid divide by zero problem. The general formula for finding the selection process is (Table 1):

Table 1: Apply the Huffman algorithm

Character	Proposed algorithm code	Probability
P	-0.0	-9.2
R	-0.2	-9.2
O	-1.1	-9.0
C	-1.1	-7.9
E	-1.2	0-6.8
D	-1.2	-5.6
U	-1.4	-4.5
R	-1.5	-3.1
E	-1.6	-1.6

$$\text{Fitness}(\text{data}) = \frac{1}{1+F_{\_obj(data)}}$$

$$\text{Fitness}[p] = \frac{1}{1+F_{\_obj[1]}} = \frac{1}{-10.972471} = -0.0911371741151104$$

$$\text{Fitness}[r] = \frac{1}{1+F_{\_obj[2]}} = \frac{1}{-8.9004874} = -0.1123533976352801$$

$$\text{Fitness}[o] = \frac{1}{1+F_{\_obj[3]}} = \frac{1}{-10.870974} = -0.0919880776092372$$

$$\text{Fitness}[c] = \frac{1}{1+F_{\_obj[4]}} = \frac{1}{-23.565322} = -0.0424352359793768$$

$$\text{Fitness}[e] = \frac{1}{1+F_{\_obj[5]}} = \frac{1}{-217917116} = -0.0458890067175816$$

$$\text{Fitness}[d] = \frac{1}{1+F_{\_obj[6]}} = \frac{1}{-22.930033} = -0.0436109272062539$$

$$\text{Fitness}[u] = \frac{1}{1+F_{\_obj[7]}} = \frac{1}{-5.9549732} = -0.1679268682519008$$

$$\text{Fitness}[r] = \frac{1}{1+F_{\_obj[8]}} = \frac{1}{-8.9004874} = -0.1123533976352801$$

$$\text{Fitness}[e] = \frac{1}{1+F_{\_obj[9]}} = \frac{1}{-21.7917116} = -0.0458890067175816$$

Total = 0.7535830918676020

The chromosome probability is formulated by:

$$p(p) = \frac{-0.0911371741151104}{-0.7535830918676020} = 0.120938454031984$$

$$p(r) = \frac{-0.1123533976352801}{-0.7535830918676020} = 0.149092248549307$$

$$p(o) = \frac{0.0919880776092372}{-0.7535830918676020} = 0.1220675975906830$$

$$p(c) = \frac{-0.0424352359793768}{-0.7535830918676020} = 0.0563112899391223$$

$$p(e) = \frac{-0.0458890067175816}{-0.7535830918676020} = 0.056311289939122300$$

$$p(d) = \frac{-0.0436109272062539}{-0.7535830918676020} = 0.0578714247664091$$

$$p(u) = \frac{-0.1679268682519008}{-0.7535830918676020} = 0.22283789281382$$

$$p(r) = \frac{-0.1123533976352801}{-0.7535830918676020} = 0.149092248549307$$

$$p(e) = \frac{-0.0458890067175816}{-0.7535830918676020} = 0.056311289939122300$$

Cumulative probability:

$$C[p] = -0.0911371741151104$$
$$C[r] = -0.203490571750391$$
$$C[o] = -0.12337134784276$$
$$C[c] = -1.16580658382214$$
$$C[e] = -1.21169559053972$$
$$C[d] = -1.25530651774597$$
$$C[u] = -1.42323338599787$$
$$C[r] = -1.5355867863315$$
$$C[e] = -1.58147579035074$$

Then generate the random number:

$$R[p] = -0.0$$
$$R[r] = -0.2$$
$$R[o] = -1.1$$
$$R[c] = -1.1$$
$$R[e] = -1.2$$
$$R[d] = -1.2$$
$$R[u] = -1.4$$
$$R[r] = -1.5$$
$$R[e] = -1.6$$

The final code for the procedure is -9.2. In the 9.2 to apply the SDC code. The SDC code for 9.2 is [I, b]. The compressed data of procedure are [I, d]. Using this proposed compression techniques to achieve the 80% of compression ratio (Fig. 1).

Calculate entropy rate: To calculate the expected amount of information has been sent for every transmission to use entropy rate calculation. The entropy rate calculation was used by Reeves (1995):

$$H = \sum_{i=1}^{m} P_i \log P_i \frac{\text{Bits}}{\text{Charactor}}$$

The model used by Shannon to find the entropy rate is:

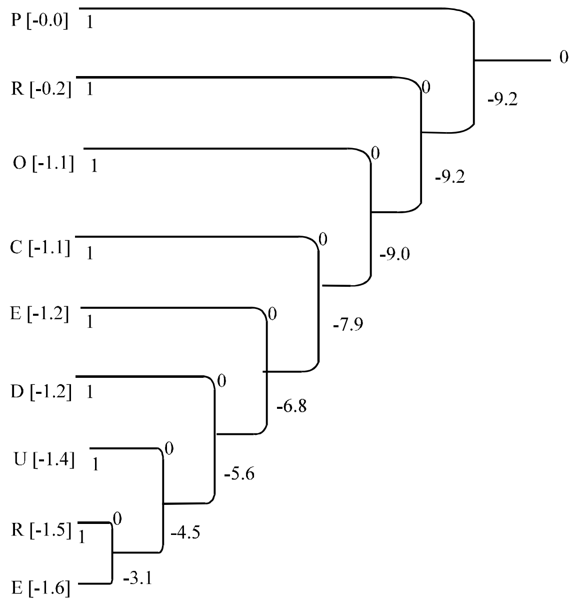

Fig. 1: Code for the procedure

$$H = \lim_{n \to con} \frac{1}{n}^{\sum P(En)\log_2 \frac{P(Bn)bits}{Charactor}}$$

Entropy rate for the input value " PROCUDURE" is:

$$H(P) = \sum_{i=1}^{1} 0.0 \log 0.0 = 0$$

$$H(R) = \sum_{i=2}^{2} 0.2 \log 0.2 = 0.6$$

$$H(O) = \sum_{i=3}^{3} 1.1 \log\log 1.1 = 0.04$$

$$H(C) = \sum_{i=4}^{4} 1.1 \log\log 1.1 = 0.04$$

$$H(E) = \sum_{i=5}^{5} 1.2 \log\log 1.2 = 0.07$$

$$H(D) = \sum_{i=6}^{6} 1.2 \log\log 1.2 = 0.07$$

$$H(U) = \sum_{i=7}^{7} 1.4 \log\log 1.4 =$$

$$H(R) = \sum_{i=8}^{8} 1.5 \log\log 1.5 = 0.17$$

$$H(E) = \sum_{i=9}^{9} 1.6 \log 1.6 = 0.20$$

Performance and measurement: Data compression techniques can be measured by the following factors (Table 2 and Fig. 2):

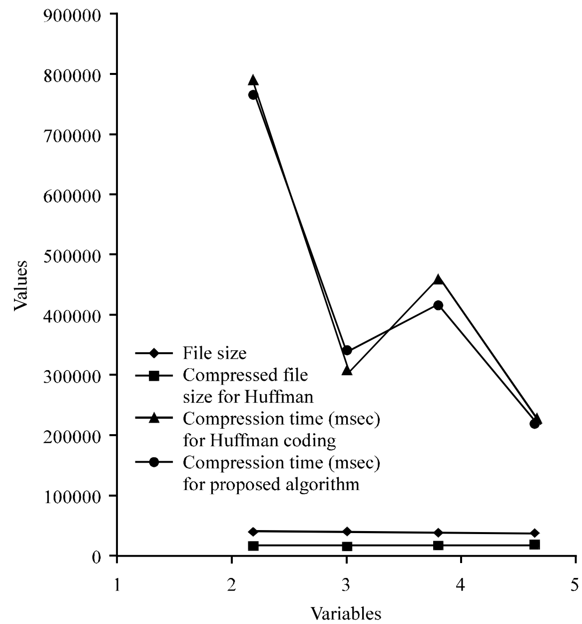

Fig. 2: Comparison of compression time of Huffman algorithm and proposed algorithm (SSG)

Table 2: Comparison of compression time of Huffman algorithm and proposed algorithm (SSG)

File size	Compressed file size for Huffman coding	Compression time (msec) for Huffman coding	Compression time (msec) for proposed algorithm (SSG)
25.000	16.100	774489	766789
21.252	11.217	308925	339876
18.510	10521	436250	417678
19.620	11610	238912	219876

$$\text{Compression ratio} = \frac{\text{Uncompressed size}}{\text{Compressed size}}$$

$$\text{Saving space} = 1 - \frac{\text{Compressed size}}{\text{Uncompressed size}}$$

Compression time =
The time taken by the algorithm to compress file

CONCLUSION

In the recent year sensor is one of the main factors in this world. But the main drawback of sensor has limited power and limited capacity. In this study, to improve the sensor efficiency through data compression techniques. Using data compression techniques to save the memory space and transmission time. In this study, to propose Sequence Statistical Code Based Data Compression Model using Genetic algorithm (SSG). Using this algorithm to get better data compression ratio. And also, the computation processes is easy when compared with Huffman coding.

REFERENCES

Akyildiz, I.F. and I.H. Kasimoglu, 2004. Wireless sensor and actor? networks: Research challenges. Ad Hoc Networks, 2: 351-367.

Anastasi, G., M. Conti, M. di Francesco and A. Passarella, 2009. Energy conservation in wireless sensor networks: A survey. Ad Hoc Networks, 7: 537-568.

Bahadili, A.H., 2008. A novel lossless data compression scheme based on the error correcting Hamming codes. Comput. Math. Appl., 56: 143-150.

Bashir, L.Z., 2015. Solve simple linear equation using evolutionary algorithm. World Sci. News, 19: 148-167.

Basu, K., C. Murthy and P. Mishra, 2013. Bitmask aware compression of NISC control words. Integr. VLSI J., 46: 131-141.

Blanco, A., M. Delgado and M.C. Pegalajar, 2000. A Genetic algorithm to obtain the optimal recurrent neural network. Intl. J. Approximate Reasoning, 23: 67-83.

Hanib, N.B.M., F. Choong, M.B.I. Reaz, N. Kamal and T. Badal, 2017. Bit swapping linear feedback shift register for low power application using 130 nm complementary metal oxide semiconductor technology. Intl. J. Eng. Trans. B. Appl., 30: 1126-1133.

Huang, G., Y. Chen, Y. Zhou, P. Liu and S. Liu, 2013. A new data processing algorithm based on model fitting for wireless sensor networks. Procedia Comput. Sci., 17: 1154-1161.

Kuo, R.J., C.H. Chen and Y.C. Hwang, 2001. An intelligent stock trading decision support system through integration of Genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets Syst., 118: 21-45.

Kwok, Y.K. and I. Ahmad, 1997. Efficient scheduling of arbitrary task graphs to multiprocessors using a parallel Genetic algorithm. J. Parallel Dist. Comp., 47: 58-77.

Lecuire, V., C. Duran-Faundez and N. Krommenacker, 2007. Energy-efficient transmission of wavelet-based images in wireless sensor networks. EURASIP. J. Image Video Process., 2007: 1-11.

Moaref, A. and V. Sattari-Naeini, 2017. Fuzzy-rough information gain ratio approach to filter-wrapper feature selection. Intl. J. Eng. Trans. C. Aspects, 30: 1326-1333.

Reeves, C.R., 1995. A Genetic algorithm for flowshop sequencing. Comput. Operat. Res., 22: 5-13.

Tang, C. and C.S. Raghavendra, 2004. Compression Techniques for Wireless Sensor Networks. In: Wireless Sensor Networks, Raghavendra, C.S., K.M. Sivalingam and T. Znati (Eds.). Springer, Boston, Massachusetts, ISBN:978-0-387-35269-5, pp: 207-231.

Wang, L., H.J. Siegel, V.P. Roychowdhury and A.A. Maciejewski, 1997. Task matching and scheduling in heterogeneous computing environments using a Genetic-algorithm-based approach. J. Parallel Distrib. Comput., 47: 8-22.

Yokota, T., M. Gen and Y.X. Li, 1996. Genetic algorithm for non-linear mixed integer programming problems and its applications. Comput. Ind. Eng., 30: 905-917.