ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.2, Pages : 21 - 25 (2014)
Special Issue of ICACE 2014 - Held on March 10, 2014 in Hotel Sandesh The Prince, Mysore, India

GA as a Data Optimization Tool for Predictive Analytics

Chandra.J[1], Dr.Nachamai.M[2] ,Dr.Anitha.S.Pillai[3]
[1]Assistant Professor, Department of computer Science, Christ University, Bangalore,India,
chandra.j@christunivesity.in
[2]Assistant Professor, Department of computer Science, Christ University, Bangalore, India,
nachamai.m@christuniversity.in
[3]Dr Anitha.S.Pillai , Professor & Head, Department of Computer Applications, Hindustan University,Chennai,India, anithasp@hindustanuniv.ac.in

Abstract*:* For any application, data used for predictive analysis has to be pre-processed .Pre-processing of data involves various steps in action based on quality of data and the nature of the application. Prior to pre-processing a general check is on missing data. The data is a dominant entity (for data analysis) when an analysis is made on it. The crux of the paper is to identify the optimality in data. This work is novel in its idea in identifying optimality in data before starting any pre-processing task. Irrespective of domain and application for which the data is to be used; the methodology approached here can be adopted. For optimality checking GA is used .The proposed method has yielded satisfactory results when applied on the different data sets. The vulnerability of the method has been proved as an attempt for two different data sets.

Key words: Genetic Algorithm, Phenotype, Genotype, Mutation, Crossover, Fitness function.

INTRODUCTION

The usage of the genetic algorithm is found in various fields such as economics, physics, mathematics, biology and many other fields. The basic idea of genetic algorithm begins with a general population of arbitrarily created individuals and it is an iterative method in which the residents in a monotonous are referred to as an invention. For every creation, the value of fitness method for each entity in the resident is measured. An extra healthy entity is probably picked from the present residents, and to generate a new population, all individual's entities are modified or rearranged and probably haphazardly mutated to form a recent invention. Normally, the genetic algorithm stops once it reached the highest quantity of creations have been shaped. The GA needs the inherited version of the solution area and the fitness function needs to evaluate the data set. The paradigm demonstration of each applicant outcome is as a group of string bits. Group of previous types and forms can be used in efficiently the same way. The major assets which make these inherited formations are correct and those portions are simply associated with their permanent size, which is used to assist the straight forward, intersect operations. Once the inherited version and the fitness functions are clear, a genetic algorithm further assign a residents of solutions and subsequently to recover it during iterative function of the crossover, mutation, inversion and choice/selection operators. The preliminary residents are shaped from an arbitrary mixture of solutions which are equivalent to chromosomes It involves the illustration of an individual a possible result or opinion or assumption in the form of its genetic formation a data structure depicting a sequence of genes called genome .The first process needs to do is the selection of the data set for predictive analytics. Once the data set is selected, it is very important to check whether the data is optimal or not. In GA, as an alternate of creating a preliminary residents having genomes of arbitrary rate with a chaotic genome is created.

RELATED WORK

Pongcharoen, P Khadwilard.A and Klakankhai.A have proposed the integrated action that ensures the realistic solutions such as the genome initialization system, reproduction and transformation methods [6]. The author's have evaluated the algorithm with the aid of three sizes of benchmarking dataset of logistic sequence network that are traditionally faced by most worldwide manufacturing companies. Abdelmaguid T.F has developed an inherited illustration and has used an arbitrary edition of a previously urbanized building heuristic in order to produce the preliminary chaotic residents for centralized stock delivery [7]. Radhakrishnan.P has released an innovative and proficient progress that facilitates on GA in regulation to disjointedly build a selection that the most feasible surplus store stage and deficiency level essential for optimal stock in the SCM and the whole prize of supply chain is controlled[8][10].

METHODOLOGY

This work aims to identify the optimal data for stock market and cervical cancer. The data from the two different domains are tested with GA to estimate the outstanding prevailing in the data. The pioneer attempt that is the pinnacle of the work is using GA to check for optimality in the data. The flow of the proposed methodology is shown in the fig.1.

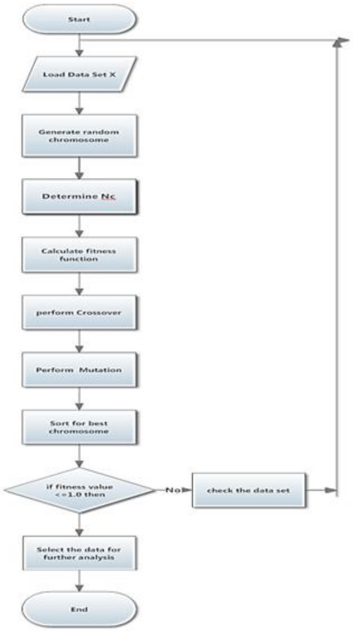

Start
Load Data Set X
Generate random chromosome
Determine Nc
Calculate fitness function
perform Crossover
Perform Mutation
Sort for best chromosome
if fitness value <=1.0 then
No
check the data set
Select the data for further analysis
End

Fig. 1: Flow of the system

There are various genetic operators are used to check the optimality of the data set for data analytics.

1. Initialization: In the beginning, all entity results are arbitrarily created to form preliminary residents. The resident's volume is based on the standard world of the problem, but usually contains a number of all probable results. Usually, the general residents are created arbitrarily, allowing the complete collection of probable results. The value of the function is measured with the given equation. The function,

$$f(x) = log\left(1 - \frac{Np}{Nc}\right), k = 1, 2, 3, \ldots, m \qquad (1)$$

Where the Nc is the quantity of counts that occurs during the phase and where m is the sum of gnome for which the healthy process is considered. The *Np* is the sum of individual values obtained. In each genome creation, the fitness method is carried out to the next level and the created genomes are arranged on the basis of fitness function. Where the Nc is the quantity of counts that occurs during the phase and where m is the sum of gnome for which the healthy process is considered. The *Np* is the sum of individual values obtained. In each genome creation, the fitness method is carried out to the next level and the created genomes are arranged on the basis of fitness function.

2. Chromosomes Representation: A preliminary residents is created from a jumbled collection of result. It involves the representation of an entity in the form of its genetic structure. At each point of the search process, a discovery of individuals is maintained.

3. Mutation: For mutation, the freshly formed genomes are given from the reproduction process. By doing the mutation, newest genome can be generated. The mutation process is completed by an arbitrary creation of two points and then doing exchanges between both the genes .This is the most basic way to alter a result for the next creation. Operators from the local search techniques may be used slightly to handle with the result for introducing new random information. The meaning of the process is about by arbitrarily changing single or extra digits in the gnome demonstrating an creature. In binary coding, the mutation may simply mean changing a 1 to a 0 and vice versa.

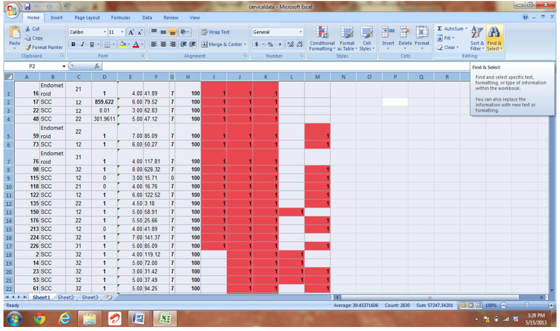
4. Crossover: Crossover or reproduction is as in biological systems, candidate solutions combine to produce children in each genetic algorithm iteration is called a generation. From the generation of parents and children, the survival of the fitness to become applicant results in the subsequent creation. The crossover is randomly picking one or more pairs of individuals as parents and randomly swapping genes of the parents. This result combines to form children for the next generation. Sometimes they pass worst information, but if the recombination is done in grouping with a powerful selection method. Recombination may be performed using different methods such as one-point crossover, N-point crossover and homogeneous recombination.

5. Termination: The genomes creation method is repeated till it reaches the stopping circumstance has been reached. In general, the stopping situations satisfy the least criteria. The uppermost position strengthens the final result or it has reached a maximum level, but the straight forward repetitions never produce an improved result. For every point of generation, the exploration method, a production of individual is maintained. The preliminary residents ideally have various persons. This termination process is essential because the individuals study from each other.

6. Fitness Method: This method is used to review the values of the genome. To evaluate how secure a given design pattern is to achieve the specified objective. In meticulously, in the fields of GP and GA, each planned result is showed as a sequence of information (also called as a genome). Behind every surrounding of checking, the preparation is to remove the N most poorly created result, and to sort N newly created from the most excellent design results. Every planned result, consequently, wants to be awarded a form of value, to signify how close to assemble the general pattern. The fitness method is always problem dependent. An illustration of a result may be a group of bits, where each bit forms a variety of entity, and the assessment of the bit represents whether 0 or 1. No such representation is legal, as the volume of things may go beyond the facility of the problem. This method guarantees that the progress is towards optimization by computing the robustness assessment for every creature in the given populations.

EXPERIMENTAL RESULTS

The tool "GA for Excel tool-v1.2" is used to check the optimality of any data sets. This free tool is developed by alexschrey for research work [14]. It allows the user to capture an excel data with any type of mathematical or computational instances. Optimization can be performed as maximization, minimization or the attempt to reach a target data set. Applications for this method lie in every field of effort. If the data set is in excel format, it can be optimized using this program. It has adequate facility for all kind of projects. To optimize different data, the GA is used as an optimization tool which is used to check whether the collected data is suitable for data analysis or not. So, the different datasets are optimized using GA. Two different examples are used in the implementations are shown in the table 1 and table 2.The two preliminary genomes are created at the commencement of the algorithm '500.100, 186.010, 104.791, 386.998, 469.817' and '246.099, 398.000, 44.001, 224.001, 502.999'.The last step value of 101th, the resultant genome moved towards the secondary genome after each iteration. Finally, both 100th and 101th iteration, the best genome '859.622, 0.010, 8.900, 228.099, 301.961'is obtained more than once. By comparing the achieved result from the GA with the history of records, it is found that the genetic algorithm is used as an optimization tool analyzes the best data set for data analysis.



cervicaldata - Microsoft Excel

P2

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	16	Endometroid	21	1	4.00	41.89	7	100	1	1	1		
2	17	SCC	12	859.622	6.00	79.52	7	100	1	1	1		
3	22	SCC	12	0.01	3.00	62.83	7	100	1	1	1		
4	48	SCC	22	301.9611	5.00	47.12	7	100	1	1	1		
5	59	Endometroid	22	1	7.00	85.09	7	100	1	1	1		1
6	73	SCC	12	1	6.00	50.27	7	100	1	1	1		1
7	76	Endometroid	31	1	4.00	117.81	7	100	1	1	1		
8	98	SCC	32	1	8.00	628.32	7	100	1	1	1		1
9	115	SCC	12	0	3.00	15.71	0	100	1	1	1		1
10	118	SCC	21	0	4.00	16.76	7	100	1	1	1		1
11	122	SCC	12	1	6.00	122.52	7	100	1	1	1		1
12	135	SCC	22	1	4.50	3.18	7	100	1	1	1		1
13	150	SCC	12	1	5.00	58.91	7	100	1	1	1	1	
14	176	SCC	22	1	5.50	25.66	7	100	1	1	1		1
15	213	SCC	12	0	4.00	41.89	7	100	1	1	1		1
16	224	SCC	32	1	7.00	141.37	7	100	1	1	1		
17	226	SCC	31	1	5.00	85.09	7	100	1	1	1		1
18	2	SCC	32	1	4.00	119.12	7	100		1	1	1	
19	14	SCC	32	1	5.00	72.00	7	100		1	1	1	
20	23	SCC	32	1	3.00	31.42	7	100		1	1	1	1
21	53	SCC	32	1	5.00	37.49	7	100		1	1	1	1
22	61	SCC	32	1	5.00	94.25	7	100		1	1		1

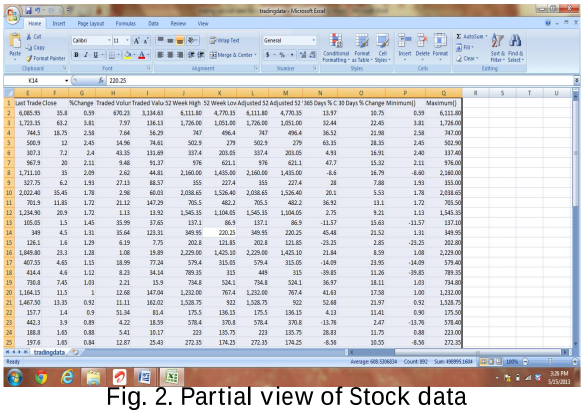
Find & Select

Find and select specific text, formatting, or type of information within the workbook.

You can also replace the information with new text or formatting.

Ready Average: 39.45371606 Count: 2830 Sum: 57247.34201

Fig:1.Partial view of cervical cancer data



tradingdata - Microsoft Excel

K14 220.25

	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Last Trade	Close	%Change	Traded Volun	Traded Valu	52 Week High	52 Week Low	Adjusted 52	Adjusted 52 '	365 Days % C	30 Days % Change	Minimum()	Maximum()
2	6,085.95	35.8	0.59	670.23	3,134.63	6,111.80	4,770.35	6,111.80	4,770.35	13.97	10.75	0.59	6,111.80
3	1,723.35	63.2	3.81	7.97	136.13	1,726.00	1,051.00	1,726.00	1,051.00	32.44	22.45	3.81	1,726.00
4	744.5	18.75	2.58	7.64	56.29	747	496.4	747	496.4	36.52	21.98	2.58	747.00
5	500.9	12	2.45	14.96	74.61	502.9	279	502.9	279	63.35	28.35	2.45	502.90
6	307.3	7.2	2.4	43.35	131.69	337.4	203.05	337.4	203.05	4.93	16.91	2.40	337.40
7	967.9	20	2.11	9.48	91.37	976	621.1	976	621.1	47.7	15.32	2.11	976.00
8	1,711.10	35	2.09	2.62	44.81	2,160.00	1,435.00	2,160.00	1,435.00	-8.6	16.79	-8.60	2,160.00
9	327.75	6.2	1.93	27.13	88.57	355	227.4	355	227.4	28	7.88	1.93	355.00
10	2,022.40	35.45	1.78	2.98	60.03	2,038.65	1,526.40	2,038.65	1,526.40	20.1	5.53	1.78	2,038.65
11	701.9	11.85	1.72	21.12	147.29	705.5	482.2	705.5	482.2	36.92	13.1	1.72	705.50
12	1,234.90	20.9	1.72	1.13	13.92	1,545.35	1,104.05	1,545.35	1,104.05	2.75	9.21	1.13	1,545.35
13	105.05	1.5	1.45	35.99	37.65	137.1	86.9	137.1	86.9	-11.57	15.63	-11.57	137.10
14	349	4.5	1.31	35.64	123.31	349.95	220.25	349.95	220.25	45.48	21.52	1.31	349.95
15	126.1	1.6	1.29	6.19	7.75	202.8	121.85	202.8	121.85	-23.25	2.85	-23.25	202.80
16	1,849.80	23.3	1.28	1.08	19.89	2,229.00	1,425.10	2,229.00	1,425.10	21.84	8.59	1.08	2,229.00
17	407.55	4.65	1.15	18.99	77.24	579.4	315.05	579.4	315.05	-14.09	23.95	-14.09	579.40
18	414.4	4.6	1.12	8.23	34.14	789.35	315	449	315	-39.85	11.26	-39.85	789.35
19	730.8	7.45	1.03	2.21	15.9	734.8	524.1	734.8	524.1	36.97	18.11	1.03	734.80
20	1,164.15	11.5	1	12.68	147.04	1,232.00	767.4	1,232.00	767.4	41.63	17.58	1.00	1,232.00
21	1,467.50	13.35	0.92	11.11	162.02	1,528.75	922	1,528.75	922	52.68	21.97	0.92	1,528.75
22	157.7	1.4	0.9	51.34	81.4	175.5	136.15	175.5	136.15	4.13	11.41	0.90	175.50
23	442.3	3.9	0.89	4.22	18.59	578.4	370.8	578.4	370.8	-13.76	2.47	-13.76	578.40
24	188.8	1.65	0.88	5.41	10.17	223	135.75	223	135.75	28.83	11.75	0.88	223.00
25	197.6	1.65	0.84	12.87	25.43	272.35	174.25	272.35	174.25	-8.56	10.55	-8.56	272.35

Ready Average: 608.5306834 Count: 892 Sum: 498995.1604

Fig. 2. Partial view of Stock data

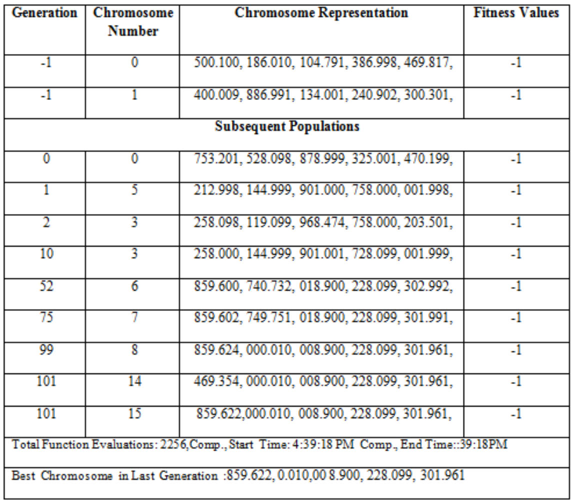
The experimental results for the stock data and cervical cancer data are shown in the following tables.

TABLE 1: EXPERIMENTAL RESULT FOR STOCK DATA

Generation	Chromosome Number	Chromosome Representation	Fitness Values
-1	0	136.001, 391.990, 44.002, 194.100, 798.523,	-1
-1	1	246.099, 398.000, 44.001, 224.001, 502.999,	-1
-1	2	136.099, 391.990, 74.001, 251.901, 502.999,	-1
Subsequent Populations			
0	0	381.124, 461.000, 557.919, 259.000, 684.000,	-1
1	0	213.999, 398.979, 510.999, 271.001, 000.010,	-1
1	15	592.999, 170.000, 100.098, 359.001, 844.993,	-1
5	14	213.999, 398.979, 510.999, 260.000, 682.010,	-1
10	8	213.995, 398.978, 352.078, 260.000, 682.901,	-1
97	7	286.955, 218.919, 141.099, 329.348, 491.491,	-1
97	8	286.955, 218.919, 141.099, 322.348, 491.491,	-1
99	13	286.957, 218.919, 141.099, 329.348, 491.491,	-1
101	10	286.955, 218.919, 141.099, 329.348, 491.491,	-1
Total Function Evaluations: 2256 Comput., Starting time: 4.01.01 Compu., End Time: 4:01:37PM			
Best Chromosome in Last Generation: 286.957, 218.919, 141.329, 329.348, 491.491			

To optimize different data in data analytics, the genetic algorithm is used as an optimization tool which is used to check whether the collected data is suitable for analysis or not. So, the different datasets are optimized using genetic algorithm. Two working examples had shown in two tables. The selection of the data set is based on the values of the fitness method. If the fitness value belongs to -1 to +1, then the selected dataset is optimal for the data analysis.

TABLE 2: EXPERIMENTAL RESULT FOR CERVICAL CANCER DATA



Generation	Chromosome Number	Chromosome Representation	Fitness Values
-1	0	500.100, 186.010, 104.791, 386.998, 469.817,	-1
-1	1	400.009, 886.991, 134.001, 240.902, 300.301,	-1
Subsequent Populations			
0	0	753.201, 528.098, 878.999, 325.001, 470.199,	-1
1	5	212.998, 144.999, 901.000, 758.000, 001.998,	-1
2	3	258.098, 119.099, 968.474, 758.000, 203.501,	-1
10	3	258.000, 144.999, 901.001, 728.099, 001.999,	-1
52	6	859.600, 740.732, 018.900, 228.099, 302.992,	-1
75	7	859.602, 749.751, 018.900, 228.099, 301.991,	-1
99	8	859.624, 000.010, 008.900, 228.099, 301.961,	-1
101	14	469.354, 000.010, 008.900, 228.099, 301.961,	-1
101	15	859.622,000.010, 008.900, 228.099, 301.961,	-1
Total Function Evaluations: 2256,Comp., Start Time: 4:39:18 PM Comp., End Time::39:18PM			
Best Chromosome in Last Generation :859.622, 0.010,00 8.900, 228.099, 301.961			

The partial views of two different data sets are given in the fig:1 , fig:2 the two preliminary genomes are created, at the start of the method. The first genome is '136.001, 391.990, 44.002, 194.100, 798.523 'and the second genome is '246.099, 398.000, 44.001, 224.001, 502.999'.The preliminary genomes are subjected for the genetic operators, reproduction and mutation .The final repetition rate of '101th', the resultant genome moved towards the resultant genome after the each iteration. Based on the fitness value, the best genome after the each iteration is selected. Finally, the execution of the 101th iteration is chosen as the best genome '286.957, 218.919, 141.329, 329.348, 491.491'was obtained. From the given experimental result, it is observed that the, the controlling resultant genome is sufficient to optimize the best data set for data analysis. Hence it is proved that the given research work is used to check the optimality of the specified data set for data analysis.

Checking, the preparation is to remove the N most poorly created result, and to sort N newly created from the most excellent design results. Every planned result, consequently, wants to be awarded a form of value, to signify how close to assemble the general pattern. The fitness method is always problem dependent. An illustration of a result may be a group of bits, where each bit forms a variety of entity, and the assessment of the bit represents whether 0 or 1. No such representation is legal, as the volume of things may go beyond the facility of the problem. This method guarantees that the progress is towards optimization by computing the robustness assessment for **d**ata set for data analysis.

CONCLUSION

The GA is used as an optimization tool for selecting any data sets in data analysis. From this research work, it is found that the GA can be considered as an optimization tool for selecting the right datasets. It is necessary to check whether the given data sets are valid not. The results elaborate the vitality requirement of optimality check on data before forwarding for any pre-processing. The methodology of using GA for optimality check has proven unrivalled and novel in its naive attempt. The proof of applying it on two different data set benchmarks the method robustness.

REFERENCES

[1] Azzini.A, A New Genetic Approach for Neural Network Design. *Ph.D. Thesis*, 2007.

[2] Azzini.A., Tettamanzi.A, A neural evolutionary approach to financial modeling, *proceeding of soft the Genetic and Evolutionary Computation Conference, GECCO- 2006*, volume No. 2, pp. 1605–1612.

[3] Azzini.A, Tettamanzi.A, Evolving Neural Networks for Static Single-Position Automated Trading, *Journal of Artificial Evolution and Applications 2008*.

[4] Azzini.A, Tettamanzi.A, Evolutionary Single-Position Automated Trading, *Proceedings of European Workshop on Evolutionary Computation in Finance and Economics, EVOFIN-2008*, pp. 1605–1612.

[5] Brabazon.A., O'Neill.M, Biologically Inspired Algorithms for Financial Modeling, *Springer*, Berlin, 2006.

[6] Pongcharoen.P, Khadwilard.A and Klakankhai.A, "Multi-matrix real-coded Genetic Algorithm for minimizing total costs in logistics chain network", *Proceedings of World Academy of Science, Engineering and Technology*, volume No. 26,pp .458-463, 2007.

[7] Abdelmaguid T.F, Dessouky M.M, A genetic algorithm approach to the integrated inventory-distribution problem, *International Journal of Production Research*- 44, pp.4445-4464, 2006.

[8] Radhakrishnan.P,Prasad.V.M and Gopalan.M.R, GeneticAlgorithm Based Inventory Optimization Analysis in Supply Chain Management,*IEEE International Advance Computing Conference,* 2009.

[9] Wang.K, Wang.Y, Applying Genetic Algorithms to Optimize the Cost of Multiple Sourcing Supply Chain Systems, *An Industry Case Study*, Volume NO. 92, pp. 355-372, 2008.

[10] Radhakrishnan.P,Prasad.V.M,Gopalan.M.R,Inventory optimization in Supply Chain Management using Genetic Algorithm *International Journal of Computer Science and Network Security* (*IJCSN)*,Volume No.1,pp.33-39,2009.

[11] Sancho Salcedo-Sanza, Jos,e-Luis Fern,and ez-Villaca nasa., Genetic programming for the prediction of insolvency in non-life insurance companies ,Volume No.5. *Computers & Operations Research* , pp.749–765,2005.

[12] Ferguson.R and Korel.B, The Chaining Approach for Software Test Data Generation,.*ACM Trans. Software Eng. and Methodology*, vol. 5, no. 1, pp. 63-86, 1996.

[13] Bingul.Z, Sekmen.A, Palaniappan.S, Sabatto.S, Genetic Algorithms Applied to Real Time Multiobjective Optimization Problems, *Proceedings of the IEEE Southeastcon* -2000 .

[14] http://www.alexschreyer.net/projects/xloptim.