International Conference on Information Engineering, Management and Security 2016 [ICIEMS]

ISBN	978-81-929866-4-7
Website	iciems.in
Received	02 – February – 2016
Article ID	ICIEMS013

VOL	01
eMail	iciems@asdf.res.in
Accepted	15 - February – 2016
eAID	ICIEMS.2016.013

Extensive Survey on Datamining Algoritms for Pattern Extraction

N Yuvaraj[1], K R SriPreethaa[2], K Kathiresan[3]
[1,2] Assistant Professor, KPR Institute of Engineering and Technology, Coimbatore. India
[3] Assistant Professor, Angel College of Engineering and Technology, Tiruppur. India

Abstract: *Volume of data available in the digital world is increasing every day at a greater speed. Due to enhancement of various technologies and new algorithms, extraction of essential data from huge volume of data is not a tough task nowadays but our goal is the extraction of patterns and knowledge from large amounts of data. Different sources are available for collecting the reviews about a product. To enhance the quality of the products and services these reviews provides different features of the products. Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set, say spam or 'ham'. Depending on definitional boundaries, modeling is synonymous with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts. In this paper we identified and discussed about three algorithms which are efficient in identifying essential patterns in the available huge volume of data.*

Keywords: *Genetic Algorithm, Shuffled frog leap algorithm, artificial neural networks.*

1. INTRODUCTION

Mining frequent patterns is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals, capacity planning and other fields. Various algorithms are useful in mining the essential data. In this paper we discussed about the working model of three algorithms namely,

- Genetic Algorithm.
- Shuffled frog leap algorithm.
- Artificial neural networks.

2. Genetic Algorithm

In recent times, computer science has seen great advancements in demands, hence implementation becomes very difficult. This situation is very apt for applying genetics and obtains optimal solutions. Genetic algorithms are search and optimization technique based on the Darwin's theory of natural evolution. The genetic algorithm is applied over the problem where the outcome is unpredictable and contains complex modules. In genetic algorithm, a population of solutions to an optimization problem is evolved towards better solutions. Each solution has a set of chromosomes (properties). Solutions from one population are taken and used to form a new population. The solutions which are selected to form new solutions called offspring are selected according to the fitness value of the solutions (chromosomes).

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

2.1 Basic Steps in the Algorithm

Step1: Random initialization populations of 'n' chromosomes are generated.
Step2: Evaluate fitness f(s) for each solution's' in the population 'n'.
Step 3: Generate a new population. Repeat the following three process until the new population is generated.

(i) Selection: Select two parent chromosomes from the population according to their fitness value. The chromosome which has the highest fitness value is more likely to be selected.
(ii) Crossover: Cross over the parent chromosome to produce the new offspring. Crossover may or may not be performed. If crossover is not performed then the new offspring is same as that of the parent chromosome.
(iii) Mutation: Mutate new offspring at each position.

Step 4: If the population in last generation is nearer to the desired solution, stop.
Step 5: Go to step 2.

2.2 Encoding

There are many ways of encoding the chromosome such as octal encoding, permutation encoding, value encoding, binary encoding, tree encoding and hexadecimal encoding. The most used way of encoding is the binary encoding. The chromosome should contain the information about the solution which it represents. Each chromosome (solution) present in the population contains a binary string. The binary string consists of 0's and 1's. Each bit in the string represents particular characteristic of problem.

Chromosome X – 11001000110100
Chromosome Y – 01110100111001

2.3 Fitness Function

Fitness function quantifies the optimality of the chromosome (solution). A fitness value is assigned to every chromosome in the population. The value is assigned based on how close it is to solve the problem.

2.4 Operators and Selection

After an initial population is generated, the algorithm evolves through the three operators, selection, crossover, mutation.

Selection is the process of selection two parent chromosomes in the population for generating new population. The parent chromosomes are selected according to their fitness. The chromosome which has the highest fitness value is more likely to be selected than the chromosome with lowest fitness value. The most common methods used for selection are Roulette wheel selection, Tournament selection, Elitism, Rank selection. In Roulette wheel selection, each chromosome in the population is assigned a slot. The chromosome with higher fitness value is assigned a larger slot and the chromosome with lower fitness value is assigned a smaller slot. The algorithm for Roulette wheel selection is simple. The weighted wheel is spinned n times (where n is the total number of solutions). When the wheel stops, the chromosome corresponds to that slot is returned. When creating new population by crossover or mutation the best chromosome may be lost. Elitism was introduced in order to retain the best chromosome at each generation. Elitism is a method which copies the best chromosome in the population to the new offspring.

2.5 Crossover

Crossover is a process in which two parent chromosome are combined to form their genetics (bits) to form new offspring which possess characteristics of both the chromosomes. Methods: Single point crossover, two point crossover, Uniform crossover. In single point crossover, one crossover point is selected, binary string from the beginning to the crossover point is copied from one point and the rest is copied from other parent chromosome.

Chromosome X	**1100010110101**
Chromosome Y	10100 01110010
Offspring 1	**11000** 01110010
Offspring 2	10100 **10110101**

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

In two point crossover, two crossover points are selected, binary string from beginning of the chromosome to the first crossover point is copied from one parent, the part from the first to the second crossover point is copied from the second parent and the rest is copied from the first parent.

Chromosome X	**1100010110101**
Chromosome Y	1010 001110010
Offspring 1	**1100** 00111 **0101**
Offspring 2	1010 **01011** 0101

In uniform crossover, bits are randomly copied from the first chromosome or from the second parent chromosome. Uniform crossover yields only one offspring.

Chromosome X	**1100110110101**
Chromosome Y	10100 01100 010
Offspring	**110**00 01**1**10 **10**0

If there is no crossover, offspring is exactly same as the parent chromosome. If there is a crossover, offspring is made from parts of parent's chromosome. If crossover probability is 100%, then all offspring is made by crossover. If it is 0%, whole new generation is made from exact copies of chromosomes from old population

2.6 Mutation

Mutation is a process by which a string is changed or inverted. Mutation probability says how often will be the parts of chromosome mutated. If there is no mutation, offspring is taken after crossover without any change. If mutation is performed, part of chromosome is changed. If mutation probability is 100%, whole chromosome is changed, if it is 0%, nothing is changed.

Parent chromosome	110**0**1011**1**001**0**
Offspring (child)	110**1**1011**0**001**1**

The most important point is that genetic algorithms supports parallelism. Genetic algorithms are used in various fields of data mining to get the optimized solutions for the better performance of the data that are required in decision making and process the accurate result. Genetic algorithm requires no knowledge about the response surface. It provides comprehensive search methodology for machine learning and optimization. Genetic algorithm has a wide scope in business. It handles large, poorly understood search spaces easily. It is easy to discover global optimal solution using Genetic algorithm.

2.7 Limitations

In many problems, Genetic algorithms may have a tendency to converge towards local optima or even arbitrary point rather than the global optimum of the problem. Finding the optimal solution to complex high-dimensional problems often requires repeated fitness function evaluations which lead to poor performance. For specific optimization problems and problem instances, other optimization algorithms may be more efficient than genetic algorithms in terms of speed of convergence. There is no effective terminator in genetic algorithm. Operating on dynamic data sets is difficult, as chromosomes begin to converge early on towards solutions which may no longer be valid for later data.

3. Shuffled frog Leap Algorithm

The shuffled frog leap algorithm is a meme tic meta-heuristic approach designed to perform an informed heuristic search to seek a global optimal solution. It is based on the evolution of memes and a global exchange of information among population. The algorithm

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

has been tested on several problems and found to be efficient in finding global solutions. The shuffled frog leap algorithm is a combination of random and deterministic approaches. In random approach, the search begins with a randomly selected population of frogs (solutions) covering the entire swamp. The population of solutions is partitioned into different subsets known as memeplexes that are permitted to evolve independently. Within each memeplex, the frogs are infected by other frog's idea, hence they experience a memetic evolution. Memetic evolution improves the quality of the meme and enhances the individual frog's performance towards a goal. During the evolution, the frogs (solutions) may change their memes using the information from the best of the entire population. After a certain number of memetic evolution time loops, the meme lexes are forced to mix and newmemeplexes are formed through a shuffling process. This shuffling enhances the quality of the memes after being infected by frogs from different regions of the swamp. The deterministic approach allows the algorithm to use response surface information effectively to guide the heuristic search.

3.1 Basic Steps in the Algorithm

Step 1: Initialization. Select x and y, where x is the number of memeplex and y is the number of frogs (solutions) in each memeplex. The size of the swamp S = xy.
Step 2: Generate a virtual population.
Step 3: Evaluate the fitness of the frogs. Sort the frogs in the order of decreasing performance.
Step 4: Partition the total population into m memeplexes.

$$X^k = f(k + m(j-1))$$

Where k varies from 0 to m, j varies from 0 to 1. For example, if m=3, rank 1 goes to memeplex 1, rank 2 goes to memeplex 2, rank 3 goesto memeplex 3, rank 4 goes to memeplex 1 etc.
Step 5: Evolution of memes in each memeplex.
Step 6: Shuffle the memeplex. If convergence criteria is satisfied then determine the best solution, stop. Otherwise, return to step 3.

3.2 Local Search

In step 5, the evolution of memeplex continues independently N times. The steps in local search for each memeplex is as follows:

Step 1: Set ix= 0 where ixcounts the number of memeplexes and will be compared with thetotal numberof x memeplexes. Set it = 0 where it counts the number of evolutionary steps.
Step 2: Set ix = ix + 1, it = it + 1.
Step 3: Determine the position of the frog in the population.
Step 4: Improve the worst frog's position.

Change in position,

$$= \text{rand}(\,) \,.\, (P_b - P_w)$$

Where P_b is the best frog's position and P_w is the worst frog's position. If this produces a better result replace worst frog.
Step 5: If step 4 cannot produce a better result, then the new position are computed for that frog is computed.
Step 6: If the new position is not better than old position, the spread of defective meme is stopped by randomly generating a new frog r at a feasible location to replace the frog.
Step 7: Upgrade the memeplex.
Step 8: If it < N, go to step 1, ix < X go to step 2. Otherwise return to shuffle memeplexes.

The Shuffled frog leap algorithm is also very suitable for parallelization. It has been used as a tool to obtain the best solutions with the least total time and cost by evaluating unlimited possible options.In this algorithm, the information gained from a change in position is immediatelyavailable to be further improved upon. This instantaneous accessibility to new information separates this approach from Genetic algorithm that requires the entire population to be modified before new insights are available.

4. Artificial Neural Networks

Artificial neural networks are a mathematical model or computational modelbased on the concept of human brain. It consists of simple processing units (artificial neurons) that communicate by sending signals to one another over a large number of interconnections. The artificial neuron transfers the incoming information on their outgoing connections to other units. It changes its structure based on external or internal information that flows through the network during the learning phase. Artificial neural networks have powerful pattern classification and pattern recognition capabilities. It can identify correlated pattern between input datasets. It can also be used to predict the outcome of the new independent input data. Artificial neural network process non-linear, complex data problems even if the data are noisy and imprecise. There are many different types of neural networks. Some of the widely used applications include classification, noise reduction and prediction.

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

4.1 Basics of Artificial Neural Networks

The working of artificial neural network has developed from the biological model of the human brain. A neural network consists of a set of connected cells. Each cell is called a neuron. The neuron receives the information from either the input cells or from other neurons and performs some kind of transformations on the input and transfers the outcome to the other neurons or output cells.

4.2 Neural Network Architectures

The two widely used artificial neural network architectures are feed forward network and recurrent networks. In feed forward network, information flows in one direction along connecting pathways, from the input layer via hidden layers to the output layer. In this network, the output of any layer does not affect that same or preceding layer.

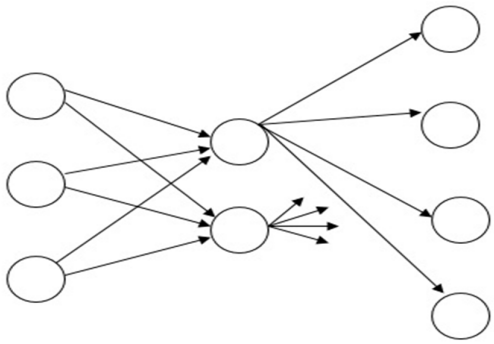

Fig: Feed forward network

In recurrent networks, there will be at least one feedback loop i.e. there could exist one layer with feedback connections. There may also be neurons with self-feedback links i.e. the output of neuron is fed back into input of itself.

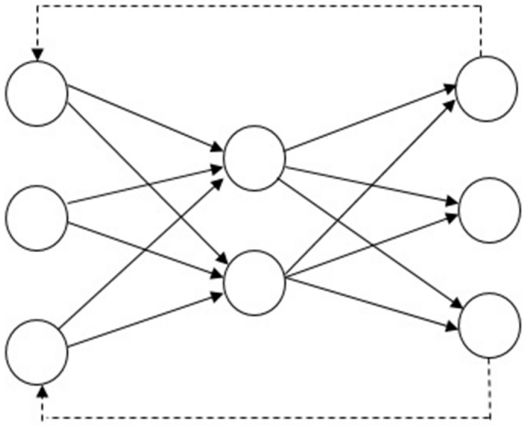

Fig: Recurrent network

4.3 Construction of ANN Model

Step 1: The input variables are selected using several variable selection procedures.
Step 2: The dataset is divided into training, testing and validation datasets. The training dataset is used to learn patterns present in the data. The learned patterns are applied on the testing data. The performance of the trained data is verified by using validation dataset.
Step 3: Define the structure of the architecture including number of hidden layers, number of hidden nodes, and number of output nodes etc.

a) Hidden layers: The hidden layer provides the network with its ability to generalize.
b) Hidden neurons: There is no certain formula for selecting optimum neurons. Some thumb rules are available for calculating hidden neurons.
c) Output nodes: Neural network with multiple output nodes will produce inferior results when compared to network with one output node.

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

d) Activation function: Activation functions are mathematical formula that determines the output of a processing node. Most commonly used activation functions are as follows:

The linear function,

$$y = x$$

The logistic function,

The hyperbolic tangent function,

$$(\exp(x) - \exp(-x)) / (\exp(x) + \exp(-x))$$

Step 4: Building the model.

Multilayer perceptron is very popular and is used more than other neural network type.

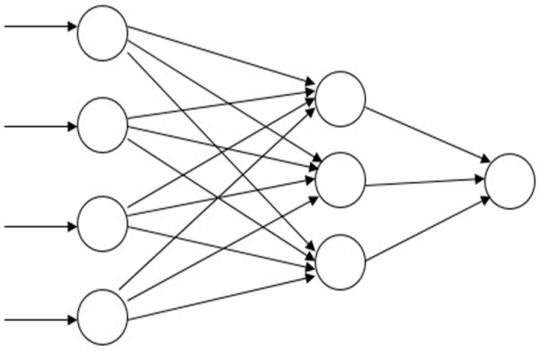

Fig: Schematic representation of neural network

4.4 Learning

Learning is a procedure that consists in estimating the parameters of neurons so that the whole network can perform a specific task. The network becomes more knowledgeable about environment after each iteration of learning process. There are three types of learning namely, supervised learning, reinforced learning, and unsupervised learning. In supervised learning, every input pattern that is used to train the network is associated with an output pattern. A comparison is made between the network computer output and the expected output, to determine the error. The error can be used to change the network parameters, which results in an improvement in performance. In unsupervised learning, there is no feedback from the environment to indicate if the outputs of the networks are correct. The network must discover features, regulations, correlations, categories in the input data automatically.

Artificial neural network with Back propagation learning algorithm is widely used in solving various classifications and forecasting problems. It can be easily implemented in parallel architectures. ANN can handle large amount of datasets and has the ability to implicitly detect complex nonlinear relationships between dependent and independent variables. Its ability to learn by example makes them very flexible and powerful.

5. Conclusion

Various algorithms discussed above has its own advantages based on the application with which its used. Based on parameters for extraction appropriate algorithm may be selected for getting efficient output.

References

1. Rashid, A., Anwer, N., Iqbal, M., & Sher, M. (2013). A survey paper: areas, techniques and challenges of opinion mining. IJCSI International Journal of Computer Science Issues, 10(2), 18-31.
2. Chandrakala, S., & Sindhu, C. (2012). Opinion Mining and sentiment classification a survey. ICTACT journal on soft computing.
3. Siqueira, H., & Barros, F. (2010). A feature extraction process for sentiment analysis of opinions on services. In Proceedings of International Workshop on Web and Text Intelligence.
4. Samsudin, N., Puteh, M., Hamdan, A. R., &Nazri, M. Z. A. (2013). Immune based feature selection for opinion mining. In Proceedings of the World Congress on Engineering (Vol. 3, pp. 3-5).
5. Pak, A., &Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.

6. Lovbjerg, M. (2002). Improving particle swarm optimization by hybridization of stochastic search heuristics and self-organized criticality. Master's thesis, Department of Computer Science, University of Aarhus.
7. Hassan, A., Abbasi, A., & Zeng, D. (2013, September). Twitter sentiment analysis: A bootstrap ensemble framework. In Social Computing (SocialCom), 2013 International Conference on (pp. 357-364). IEEE.
8. Cabanlit, M. A., &Junshean Espinosa, K. (2014, July). Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons. In Information, Intelligence, Systems and Applications, IISA 2014, the 5th International Conference on (pp. 94-97). IEEE.
9. Neethu, M. S., &Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE.
10. Liu, S., Cheng, X., & Li, F. (2015). TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets.
11. Coban, O., Ozyer, B., &Ozyer, G. T. (2015, May). Sentiment analysis for Turkish Twitter feeds. In Signal Processing and Communications Applications Conference (SIU), 2015 23th (pp. 2388-2391). IEEE.
12. Aldahawi, H., & Allen, S. M. (2013, September). Twitter mining in the oil business: A sentiment analysis approach. In Cloud and Green Computing (CGC), 2013 Third International Conference on (pp. 581-586). IEEE.
13. Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.
14. Celikyilmaz, A., Hakkani-Tür, D., & Feng, J. (2010, December). Probabilistic model-based sentiment analysis of twitter messages. In Spoken Language Technology Workshop (SLT), 2010 IEEE (pp. 79-84). IEEE.
15. C. D. Maning, P. Raghavan, and H. Schtze,An Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK,2008.
16. Ahmad, T., Doja, M. N., and Islamia, J. M. "Ranking System for Opinion Mining of Features from Review Documents", International Journal of Computer Science, Vol.9, No.4, pp.440-447, 2012.
17. Francis, L., and Flynn, M. "Text mining handbook. In Casualty Actuarial Society E-Forum", spring, pp.1-61, 2010.
18. Rivas "Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval", 2014.
19. Sarkar, C., Desikan, P., & Srivastava, J. (2013). Correlation based Feature Selection using Rank aggregation for an Improved Prediction of Potentially Preventable Events.
20. Sharma, A., &Tivari, N. (2012). A survey of association rule mining using genetic algorithm. Int J ComputApplInfTechnol, 1, 5-11.

Cite this article as: SN Yuvaraj, K R SriPreethaa, K Kathiresan. "Extensive Survey on Datamining Algoritms for Pattern Extraction". *International Conference on Information Engineering, Management and Security 2016*: 63-69. Print.