Unsupervised Feature Selection Using Evolutionary Algorithms

Ms. Aishwarya Deshpande, Ms. Sharvari Deshpande, Ms. Monika Doke, Ms. Anagha Chaudhari

Abstract— **Classification is a central problem in the fields of data mining and machine learning. Using a training set of labelled instances, the task is to build a model (classifier) that can be used to predict the class of new unlabelled instances. Data preparation is crucial to the data mining process, and its focus is to improve the fitness of the training data for the learning algorithms to produce more effective classifiers. Searching for the frequent pattern within a specific sequence has become a much needed task in the various sector. Most recent works are based on Apriori algorithm, GSP, MacroVspan etc. techniques. However, frequent pattern mining can be made more efficient. Two widely applied data preparation methods are feature selection and instance selection, which fall under the umbrella of data reduction.**

Feature selection is selecting a subset of optimal features. Feature selection is being used in high dimensional data reduction and it is being used in several applications like medical, image processing, text mining, etc. Several methods were introduced for unsupervised feature selection. Among those methods some are based on filter approach and some are based on wrapper approach.

In the existing work, unsupervised feature selection methods using Genetic Algorithm, Bat Algorithm and Ant Colony Optimization have been introduced. These methods yield better performance for unsupervised feature selection. We will use a novel method to select subset of features from unlabeled data using binary bat algorithm with sum of squared error as the fitness function.

Index Terms— **frequent pattern; Pattern Indexing; HashMap; ASCII Byte-encoding, unsupervised feature selection, Binary bat algorithm, K –means, Ant Colony Optimization (ACO), Data Mining, Classification, Data Reduction, Instance Selection.**

I. INTRODUCTION

In most of the applications we often have a very large number of features that can be used. Feature selection is a process of selecting subset of features available from large set of data [1]. The best subset of features contains least number of dimensions that contribute to the accuracy of the model by removing irrelevant and redundant features. Feature selection is used to avoid curse of dimensionality and to reduce computational burden.

Feature selection in supervised learning can be done easily since we know the class label in supervised learning it's easy for us to decide which feature we want to keep based on the class label. But class label is not present in unsupervised feature selection. Our paper gives a solution to this problem by using binary bat optimization algorithm for feature selection.

II. SYSTEM ARCHITECTURE

Data mining is the process of extracting insightful knowledge from large quantities of data either in an automated or a semi-automated fashion. Evolutionary data mining, or genetic data mining is aumbrella term for any data mining using evolutionary algorithms.

Evolutionary algorithms work by trying to emulate natural evolution. First, a random series of "rules" are set on the training dataset, which try to generalize the data into formulas.The rules are checked, and the ones that fit the data best are kept, the rules that do not fit the data are discarded. The rules that were kept are then mutated, and multiplied to create new rules.

This process iterates as necessary in order to produce a rule that matches the dataset as closely as possible.When this rule is obtained, it is then checked against the test dataset. If the rule still matches the data, then the rule is valid and is kept. If it does not match the data, then it is discarded and the process begins by selecting random rules again.

Module 1: Database

Before database can be mined for data using evolutionary algorithms, it first has to be cleaned,which means incomplete, noisy or inconsistent data should be repaired. It is imperative that this be done before the mining takes place, as it will help the algorithms produce more accurate results.

If data comes from more than one database, they can be integrated, or combined, at this point. When dealing with large datasets, it might be beneficial to also reduce the amount of data being handled. One common method of data reduction works by getting a normalized sample of data from

Ms. Aishwarya Vishwas Deshpande, Student, Department of Information Technology Savitribai Phule Pune University/Pune, India,
Ms. Sharvari Avinash Deshpande, Student, Department of Information Technology Savitribai Phule Pune University/Pune, India,
Ms. Monika Kalyan Doke, Student, Department of Information Technology Savitribai Phule Pune University/Pune, India,
Ms. Anagha Chaudhari, Assistant Professor Department of Information Technology Savitribai Phule Pune University/Pune, India,

the database, resulting in much faster, yet statistically equivalent results.

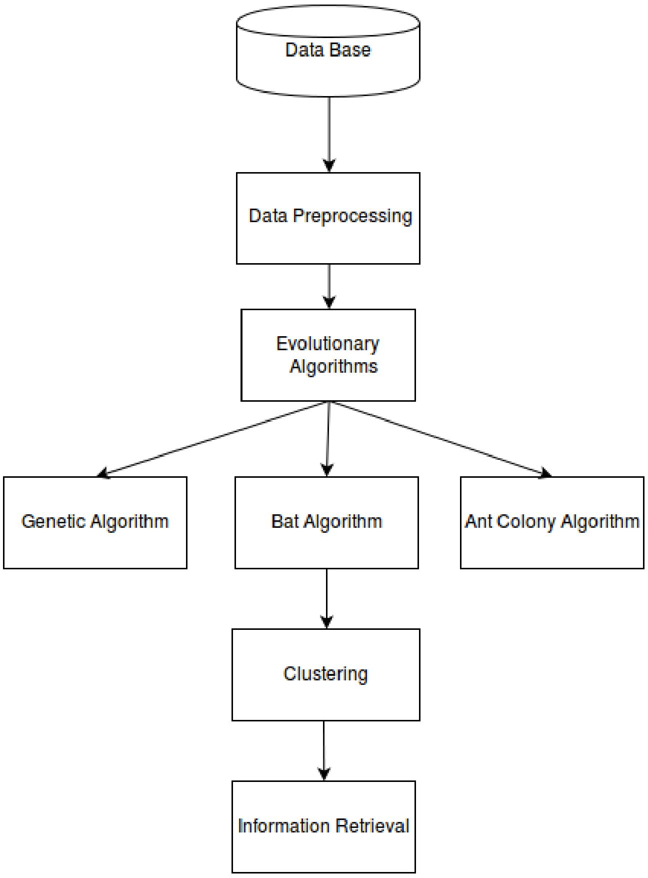


Fig.1.System Architecture

At this point, the data is split into two equal but mutually exclusive elements, a test and a training dataset. The training dataset will be used to let rules evolve which match it closely. The test dataset will then either confirm or deny these rules.

Module 2 : Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Module 3 : Evolutionary Algorithms

In artificial intelligence, an evolutionary algorithm (EA) is a subset of evolutionary computation, population based metaheuristic optimization algorithm. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation and selection.

Candidate solutionsto the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions. Evolution of the population then takes place after the repeated application of the above operators. There are three main evolutionary algorithms:

A. Genetic Algorithm

Genetic algorithm belongs to the larger class of evolutionary algorithms. Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. Searching for the frequent pattern within a specific genetic sequence has become a much needed task in the bioinformatics sector. We develop an algorithm where the subsequent nucleotide sequences of a given size (a multiple of 4 in this case, e.g. 4, 8, 12, 16 etc.) can be identified within a particular DNA sequence using an assigned numerical value. We'll also store their repetition count in a memory efficient manner by using ASCII byte-encoding, and the most repeating sequence(s) of the given length will be provided as output.

Frequent pattern matching and data mining for DNA/protein sequence analysis is being considered remarkable among the researchers. Normally data sequences are very large but in case of DNA sequence only A,C,T,G makes a nucleotide and so it is very natural that numerous combinations and permutations of A,C,T and G's will be repeated many times. Thus, the importance of recognizing the correct DNA sequence pattern is easily understandable.

There are a lot of aspects in DNA sequence that needs to be discovered for proper analysis. Reasons and causes of all diseases can be found if the data are properly analyzed and sorted. Analyzing the sequence we can find similarities and links between two sequences and it is possible to modify them accordingly. Regarding frequent pattern matching many works have been done and many works are still ongoing. However efficiently processing the vast volume of data, finding the proper sequence and important patterns are still major challenges in this field.

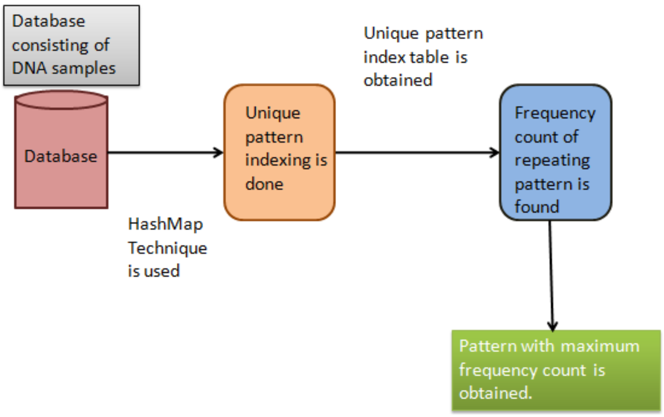

Database consisting of DNA samples
Database
Unique pattern indexing is done
HashMap Technique is used
Unique pattern index table is obtained
Frequency count of repeating pattern is found
Pattern with maximum frequency count is obtained.

Fig.2.Block Diagram of Genetic Algortihm

Unique Pattern Indexing :

Unique Pattern Indexing Algorithm indexes each unique pattern and puts it into a Hash Map.This technique will simply format the input text file with DNA sequence of a sample species and then encode it using numerical values.

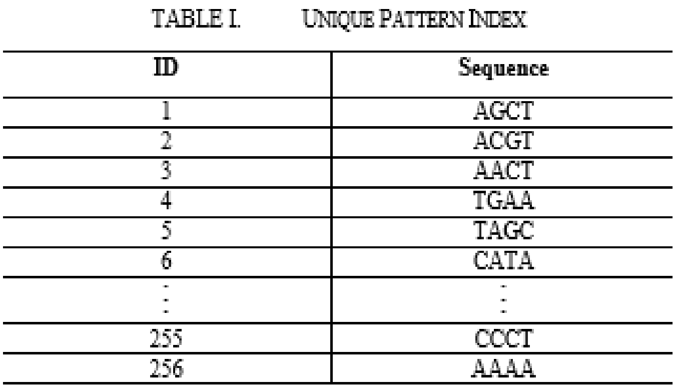


TABLE I. UNIQUE PATTERN INDEX

ID	Sequence
1	AGCT
2	ACGT
3	AACT
4	TGAA
5	TAGC
6	CATA
:	:
255	CCCT
256	AAAA

Pattern Indexing Table

Sample Dataset :

In order to build the initial database, we need some sample DNA sequences. The sample dataset are obtained from NCBI database and are of different lengths. This gives us a realistic simulation given the dataset are all obtained from different species of bacteria and viruses. This algorithm can be modified quite easily for nucleotides of other varying lengths, e.g. 3, 6, 9, 12 etc. In order to do so, we just have to change how Segment_size increments in the first algorithm.

Segment_size= i * n

where n is the length of interval.

A. Bat Algorithm

Feature selection is selecting a subset of optimal features. Feature selection is being used in high dimensional data reduction and it is being used in several applications like medical, image processing, text mining, etc. In this technique we proposed a novel method to select subset of features from unlabeled data using binary bat algorithm. Bat algorithm is a optimization algorithm developed in 2010. This Bat Algorithm is based on echolocation behaviour (location of objects by reflected sound) of micro bats with varying pulse rate of emission and loudness.

Velocity Vector :

The velocity can be updated using the following equation:

$$V_i\ (t+1) = V_i\ (t) + (X_i\ (t) - Gbest)\ F_i$$

where Vi=velocity of ith bat
Fi=frequency
Xi=position

Frequency Vector :

The frequency of bat can be updated using following equation:

$$F_i = F_{min} + (F_{max} - F_{min})\ \beta$$

where Fmin= minimum frequency
Fmax=maximum frequency

Proposed Method :

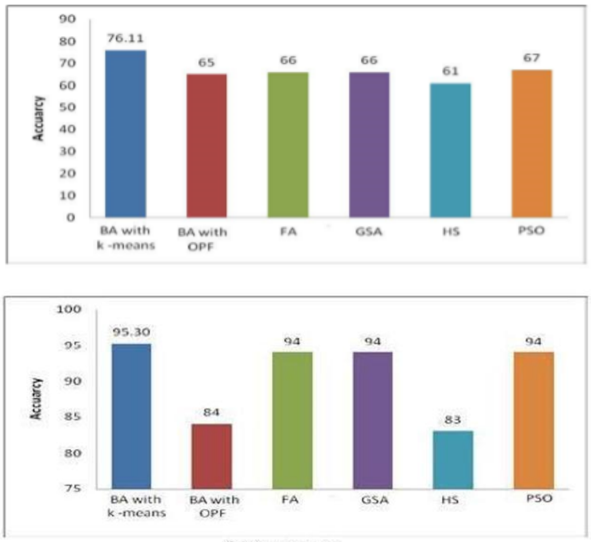
We use binary bat algorithm along with K-means clustering algorithm for feature selection. Initially the number of bats is set as half of the number of attributes in the dataset. The features with 1's are selected features and 0's are unselected features in our encoding scheme. The selected features are given to simple Kmeans clustering algorithm. The sum of squared error in K-means algorithm is used as the fitness function. Features with 1's in the final Gbest are considered as the best solution.

Sum of Squad Errors :

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} dist^2(m_i, x)$$

Experimental Results :

Recently many powerful methods for global optimizations are derived from the behavior of biological system. The features that gives highest classification accuracy is said to be the best features. The proposed method gives higher accuracy of 95.30% results for SVM GUIDE1. Figures compares the accuracy of dataset by BBA with OPF, FA, GSA, HS, and PSO against BBA with k-means. This features of Splice dataset by BBA with k-means gives highest accuracy of 76.11%. This figure gives the result for ionosphere with highest accuracy of BA with k-means of 85.75%



(a)Accuracy

Fig.3 .Experimental Results

A. Ant Colony Algorithm

Ant colony optimization (ACO) is a meta-heuristic inspired by the behaviour of real ants in finding the shortest route between two locations. We introduce ADR-Miner, a novel data reduction algorithm that utilizes ant colony optimization(ACO).

ACO has been successful in solving several types of optimization problems, in particular combinatorial optimization problems, as well as data mining problems. It is inspired mainly by the foraging behaviour observed in ant colonies to find the shortest path between a food source and the nest. In essence, the search space of a given problem is first transformed into a graph of solution components, whereby a combination of these components would present a valid candidate solution to that problem.

A population of ants then navigates this graph, selecting decision components to add to their path chosen throughout the graph in an attempt to build a candidate solution. Classically applied to combinatorial optimization problems, ACO has also been successful in tackling classification problems. Ant-Miner is the first ant-based classification algorithm, which discovers a list of classification rules.

As alluded to earlier, our algorithm adapts an ACO algorithm to perform data reduction with an emphasis on improving a classifier's predictive effectiveness. Adapting the ACO algorithm to perform data reduction involved a number of steps:

I. Construction graph.

II. Defining the overall meta-heuristic data.

III. Defining how an ant constructs a candidate solution .

IV. Defining the mechanics of evaluating the quality of such solutions and updating the pheromone trails.

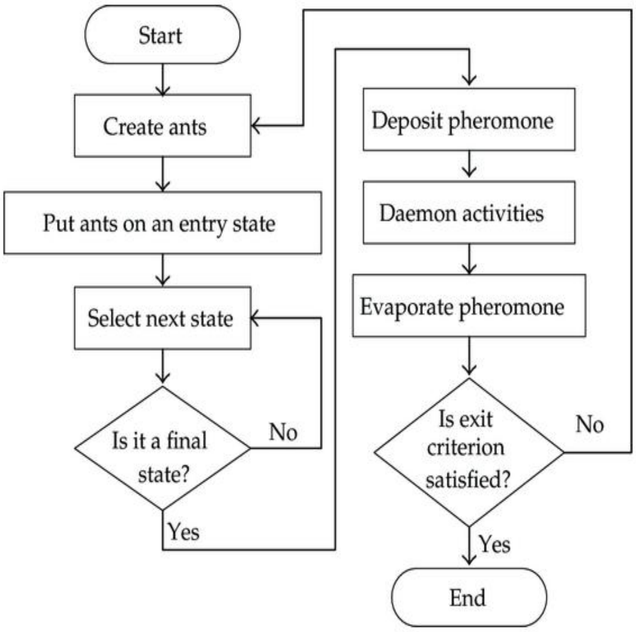


Fig.4. Flowchart

Data Reduction :

As mentioned earlier, data reduction is a vital pre-processing task for classification and its significance lies in that it removes noisy, outlier and other instances from the training data set. In addition to improving accuracy, it also reduces the size of the training set before it is presented to the machine learning algorithm. For lazy-learning algorithms, such as nearest neighbour(s), the reduced data set decreases the time needed for arriving at the class of a new instance in question.

III. CONCLUSION

This research ultimately contributes to the comprehensive choices of techniques and algorithms to be used for Information Retrieval System. A totally new approach to the Information Retrieval has been laid forward by comparing various algorithms. A brief comparison of Evolutionary algorithms like Genetic Algorithm, Bat Algorithm and Ant Colony optimization that proves Bat Algorithm can be better for effective and promising results. At the later stage Clustering techniques of totally different methodology and mechanism were examined with same data set; the very next output of clusters was treated for Information Retrieval to get better conclusions over the techniques and variety of approach combinations carried out throughout the study. All the clustering approaches like K-means clustering, Fuzzy clustering and the Hierarchical clustering gave different output for cluster plots. Thus, the overall study led to a better, effective, efficient, reliable, relevant and excellent Information System which can be user friendly and applied anywhere on textual datasets for ease of data handling, management and access through retrieval. In near future same system will be examined with various datasets for more confident solutions and conclusions.

.

REFERENCES

[1] Yu, L., Liu, H. (2003), “Feature selection for high dimensional data: a fast correlation based filter solution”. Proc. 20th Int’l Conf. Machine Learning; 856-863.

[2] H. Liu, L. Yu. (2005), “Toward integrating feature selection algorithms for classification and clustering”, IEEE Transactions on Knowledge and Data Engineering 17 (4) 491–502.

[3] M.A. Jayaram, A.G. Karegowda, A.S. Manjunath. (2010), “Feature subset selection problem using wrapper approach in supervised learning”, International Journal of Computer Applications 1 (7) 13–16.

[4] Kennedy J, Eberhart R. (1995), “Particle swarm optimization”. In: Proceedings of IEEE international conference on neural networks, pp 1942–1948.

[5] Holland J .(1975),”Adaptation in natural and artificial systems”, The University of Michigan Press, Ann Arbor, Michigan

[6] Dorigo M, Maniezzo V, Colorni. (1996),” A the ant system: optimization by a colony of cooperating agents”, IEEE Trans Syst Man Cybern B 26:29–41.

[7] Kennedy J, Eberhart RC. (1997),” A discrete binary version of the particle swarm algorithm”, IEEE.

[8] Rashedi E, Nezamabadi S, Saryazdi S (2009),”GSA: a gravitational search algorithm”. InfSci 179:2232–2248.

[9] Ramos, C., Souza, A., Chiachia, G., Falcao, A., & Papa, J. (2011),” A novel algorithm for feature selection using harmony search and its application for non-technical losses detection”, Computers & Electrical Engineering, 37, 886–894.

[10] H. Hannah Inbarani, P.K. Nizar Banu, S. Andrews.(2012), “Unsupervised hybrid PSO - quick reduct approach for feature reduction”, Proceedings of International conference on Recent Trends in Information Technology (2012) 11–16.

[11] S. Bai, S. X. Bai, “The Maximal Frequent Pattern Mining of DNA Sequence,” GrC, pp 23-26, 2009.

[12] S. F. Zerin, B. S. Jeong, “A Fast Contiguous Sequential Pattern Mining Technique in DNA Data Sequences Using Position Information,” Department of Computer Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, 446-701, Korea. Date of Web Publication 12-Dec-2011.

[13] T. H. Kang, J. S. Yoo and H. Y. Kim, “Mining frequent contiguous sequence patterns in biological sequences,” in proceeding of the 7th IEEE International Conference on Bioinformatics and Bioengineering, pp. 723-8, 2007.

[14] J. Pan, P. Wang, W. Wang, B. Shi and G. Yang, "Efficient algorithms for mining maximal frequent concatenate sequences in biological datasets", in Proceedings of the 5th International Conference on Computer and Information Technology(CIT), pp. 98-104, 2005.

[15] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules.” In Proc. 1994 Int. Conf. Very Large Databases (VLDB?94), pages 487–499, Santiago, Chile, Sept. 1994.

[16] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Addison Wesley, 2005.

[17] C. M. Bishop, Pattern Recognition and Machine Learning . Berlin, Heidelberg: Springer, 2007.

[18] K. M. Salama and A. M. Abdelbar, “A novel ant colony algorithm for building neural network topologies,” in 9th International Conference on Swarm Intelligence (ANTS’14), ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2014, vol. 8667, pp. 1–12