# Data Mining Evolutionary Learning (DMEL) using H base

Miss. Mansi Shah, Ms. Seema Kolkur

*#M.E. Computer, Department of Computer Engineering, Mumbai University, India*

**Assistant Professor, Department of Computer Engineering, Mumbai University, India*

***Abstract***— *In current market scenarios, telecom companies are quite competitive and look forward to have lion's share in the market by winning new and withholding existing customers. Customers who are lost to competitor are known as Churned customers and can be retain by adopting Churn prevention model. For a given dataset, this model predicts the list of customers to be churned in future enabling the respective authorities to take action accordingly. However in telecom, the results of algorithms suffer due to disproportion nature and vast size of datasets. In this paper, Genetic Programming (GP) based approach for modelling the challenging problem of churn prediction is incorporated in HBASE. A data mining algorithm, Data Mining Evolutionary Learnings (DMEL), handles a classification problem which helps to meet accuracy of prediction. As data in telecom industry is going to increase so to make the classification process fast DMEL algorithm is incorporated in HBase. For competitive telecom industry, churn prediction approach would be significantly beneficial.*

**Keywords** — *telecom industry, churn prediction, genetic algorithms, DMEL, HBase.*

## I. Introduction

Indian telecom is one of the biggest economies posting a robust growth rate more than 35 percent over past decade in terms of subscribers. Increasing innovation and deregulation by government has created a competitive market, empowering people to choose from plethora of offers and services. Every business looks forward to build and maintain loyal customer base, however, this fierce competition have resulted in risk of customers switching to competitor. In such aggressive and open market conditions, it is crucial to know beforehand which customers are going to switch to a competitor, since acquiring new customers is expensive than retaining existing ones. When a customer switches to competitor that customer is known as lost customer or churn customer. Churn prevention technique is quite economical and gives an edge over competitors. Customer churn is broadly divided into categories a) voluntary and b) involuntary churn. Voluntary churn, service contract is terminated by customer, in case of involuntary churn, customer fraud, non-payment or underutilized subscription service is disconnected by the company [1].

Churn is an expensive venture, if not managed carefully, could bring to the company its yield. Income and expenses related to the loss of customers, includes losses, costs and regain customer retention, advertising costs, organizational chaos and planning, and budgeting. Furthermore, previous studies have shown that the cost of retaining the existing customer is lower than the cost acquiring new customers. Therefore, it makes business sense for players in the industry to identify those who may leave the company in advance and to develop intervention strategies in the fight to retain as many customers as subscribers.

Customer churn prediction in data mining is currently a relevant topic, has been used in banking, mobile telecommunications, life insurance, and other areas. In fact, all the companies who are dealing with long-term customers can take advantage of customer churn prediction method. The goal is to differentiate as much as possible from the non-stirrer agitator [2].

Data classification is an important issue in mining research. Many algorithms for classification model mine's large data sets; they have proven to be very effective. However, determining the likelihood of each category, a lot of people are not in line with this purpose [1]. Many data mining algorithms, such as decision tree algorithms (for example, the ship, public rainforest, SLIQ, sprint) generates rules, used to find records of unknown class members. However, when the extended decision tree algorithm determines, based on the probability associated with this classification, it is possible that some of the leaves have a similar class probabilities in the decision tree. Decision tree algorithm, logit regression and neural networks, and other classification techniques can accurately determine the different probabilities to predict its possibilities. However, compared with the decision tree algorithm these algorithms are explicit in expression of symbolic, easily understood form (for example, if - then rules) in discovery mode. Because of the limitations of these prior art, p a new algorithm, Data Mining Evolutionary Learning (DMEL). DMEL algorithm has the following features.

1. Instead of randomly generated initial population, consisting of a set of first-order rule *l*, generates an initial population using induction technology. Based on these rules, the rule with a higher order is obtained using initial population.
2. While determining the rules of interest, DMEL does not require the subjective interest instead of it uses objective interestingness measure.
3. In the evaluation of chromosome fitness, DMEL uses a function that can properly determine its encoding rules. This function is defined in terms of probability.
4. It can estimate prediction of likelihood made.
5. DMEL handles missing values effectively [1].

.

## II. Literature Review

### A. HBase

Apache HBase, a NoSQL database, can host tables with huge number of records because of Hadoop Distributed File System (HDFS) and Hadoop's MapReduce programming model. These features makes it a robust database that provides the flexibility to generate result for real time individual record query as well as result for batch processing of massive data [3].

### B. Bayesian Discretization

Numeric variables can be continuous or discrete. A continuous variable takes an infinite number of possible values within a range or an interval. A discrete variable is one which takes a countable number of distinct values [4]. Discretization converts the variable from discrete to continuous or vice versa. This process creates set of contiguous intervals. The set of intervals or the set of cut points generated by a discretization method is called discretization.

Discretization has several advantages. It can be applied to expand the data set, because some classification algorithms cannot handle continuous attributes. In addition, pre-processing steps are necessarily required for classification method that requires discrete data. Discretization helps to achieve this as well as help to increase accuracy of some of the classifier, increase the speed of classification, especially in high-dimensional data, and provide a better model. Main disadvantage of discretization is it leads to loss of information resulting in reduction of performance of classifier.

Discretization can be classified as supervised or unsupervised. The information about the target variable is not used in unsupervised discretization process while supervised methods do. Compare to unsupervised methods, supervision methods tend to be more complex, and are often having excellent performance classifier. FI common method is an example of supervised methods.

### C. DMEL

An evolutionary learning approach, Binary attributes are easily mined by DMEL algorithm from large dataset without any user-defined thresholds. However, Discretization algorithm is used to convert quantitative attributes to categorical attributes which is used by DMEL algorithm [1]. Number of rules defines number of condition in the antecedent of a rule. For e.g. if antecedent of rule has one condition than it is said as one-condition rule i.e. first-order rule and so on. DMEL discovers rules by an iterative process. Probabilistic induction technique is used to generate first-order rule. Using this rules higher order rules are generated in each iteration [1].

## III. Concept of the Proposed System

### A. Proposed System Perspective

The proposed system aims to simplify the job handled by businessman whether it is business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting. By merely giving the mobile consumer data records as input, user can get the desired consumer behaviour pattern. This system focuses on churn prediction for telecom industry. The system is built using Apache HBase [2].

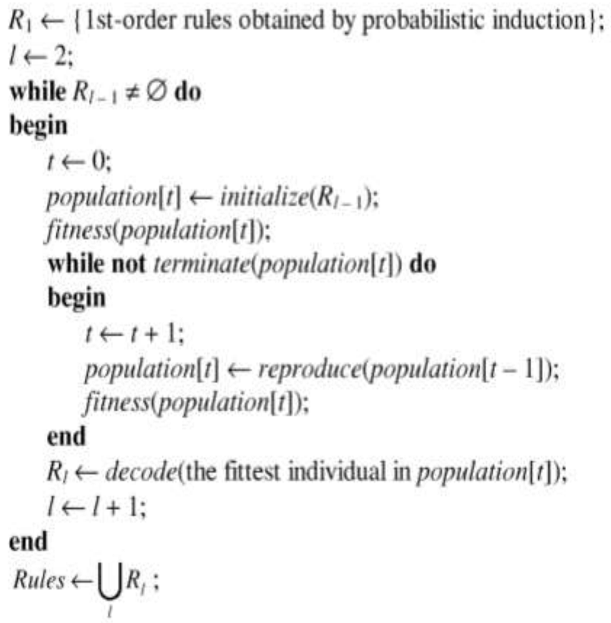
### B. Proposed System Features

1. Ability to provide "just-in-time" information: Provide managers with the information they need to make effective decisions about an organization's strategic directions [2].
2. Diverse pattern Generation: Diverse hidden patterns can be generated from input data which otherwise has no meaning and is unordered [2].
3. Reduced restrictions on building query: Because of HBase the restriction on writing query is reduced [2].
4. Complex Calculations: User may have unrelated data which is difficult to analyze and process. The complex calculation involved in mining data is resolved by the combination of HBase and DMEL algorithm [2].

5. Ability to model real business problems and more efficient use of people resources: By providing the ability to model real business problems and a more efficient use of people resources, this project enables the organization as a whole to respond more quickly to market demands. Market responsiveness, in turn, often yields improved revenue and profitability [2].
6. Data storage reliability: Storing data in HBase cluster helps to have replica of each and every information and in the event of failure we can use replica and no data loss [2].

## IV. Proposed Methodology

The proposed methodology uses evolutionary algorithms, the so-called genetics-based machine learning algorithm, DMEL and HBase in order to provide a state of the art, evolutionary rule based systems is applied to classification tasks. Every individual is represented by a different rule which is evolved until convergence. The tasks of regression, clustering and association discovery are the promising tasks that are solved by GP. GP also offers features like flexibility and distinctiveness which makes it excellent technique for classification. Here we have considered the problem churn prediction for telecom industry.Fig.1 depicts the DMEL algorithm [5].



```
$R_1 \leftarrow$ {1st-order rules obtained by probabilistic induction};
$l \leftarrow 2$;
while $R_{l-1} \neq \emptyset$ do
begin
    $t \leftarrow 0$;
    $population[t] \leftarrow initialize(R_{l-1})$;
    $fitness(population[t])$;
    while not $terminate(population[t])$ do
    begin
        $t \leftarrow t + 1$;
        $population[t] \leftarrow reproduce(population[t-1])$;
        $fitness(population[t])$;
    end
    $R_l \leftarrow decode$(the fittest individual in $population[t]$);
    $l \leftarrow l + 1$;
end
$Rules \leftarrow \bigcup_l R_l$;
```

Fig.1 DMEL Algorithm

### A. Encoding Rules in the Chromosomes

The complete sets of rules are encoded in single chromosome i.e. each gene in each single chromosome gets encoded as a single rule. The only part that is not encoded is the consequent and the uncertainty because it should not be determined by any chance. Only while computing fitness of chromosome the consequent and uncertainty is determined in DMEL [1]

### B. Generating First-Order Rules

DMEL consists of a set of first-order rules at the beginning of an evolutionary process. When compared with the initial population generated randomly, initial population generated heuristically improves the speed of convergence and find the better solutions. Based on these findings, DMEL first identifies a set of first-order rules, and place it in the initial populations. Initial population containing\consisting first order rules are generated in negligible amount of time using interestingness measures and weight of evidence measures first order rules are achieved. To do so, APACS probabilistic induction technic is used. From all possible attribute-value pairs, APACS, identifies those who have some kind of relationship, even if the database is noisy and contains many missing values. Eqn.1 gives the equation of weight of evidence in terms of probability [1].

$$\hat{w}_{a_{ip}a_{jq}} = \log \frac{\hat{\Pr}(A_j = a_{jq} | A_i = a_{ip})}{\hat{\Pr}(A_j = a_{jq} | A_i \neq a_{ip})}.$$

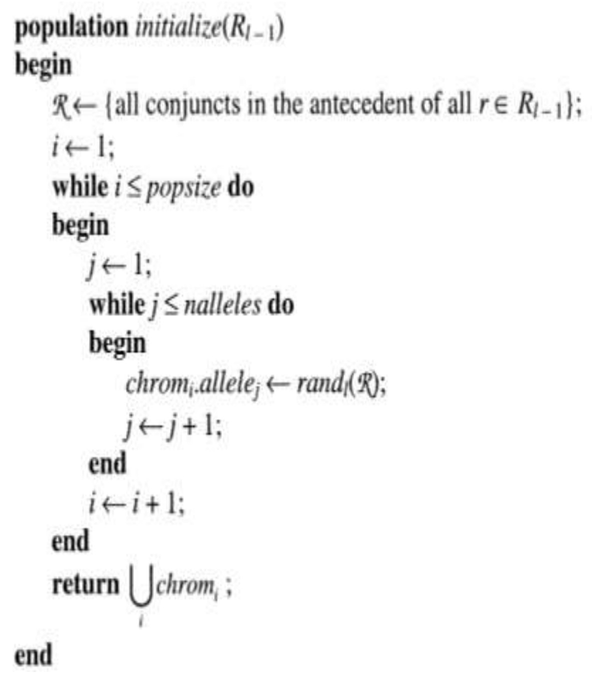
Eqn.1 Equation of weight of evidence

Where, A=set of attributes.

$a_{ip}$ and $a_{jp}$ = allele represents the rule.

### C. Initialization of Populations

Good initial population can improve the speed of the evolutionary process and makes it easier to find the best solution. DMEL does not produce its initial population complete randomly; instead, it uses a heuristic. According to this heuristic, DMEL combines randomly (*l*-1)th order rules generated in previous iteration to generate *l*th order rules. Fig.2 depicts that, how the initialize function will work. Initialize function helps to select the initial population to generate rules [1].



```
population initialize($R_{l-1}$)
begin
    $\mathcal{R} \leftarrow$ {all conjuncts in the antecedent of all $r \in R_{l-1}$};
    $i \leftarrow 1$;
    while $i \leq popsize$ do
    begin
        $j \leftarrow 1$;
        while $j \leq nalleles$ do
        begin
            $chrom_i.allele_j \leftarrow rand_l(\mathcal{R})$;
            $j \leftarrow j + 1$;
        end
        $i \leftarrow i + 1$;
    end
    return $\bigcup_i chrom_i$;
end
```

Fig.2 The initialize function

Where, *l* = order of rules

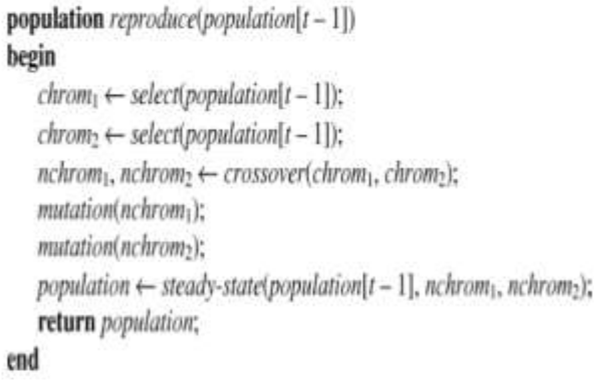
R = set of rules

*popsize* is set 30

*rand1* = random function

### D. Genetic Operators

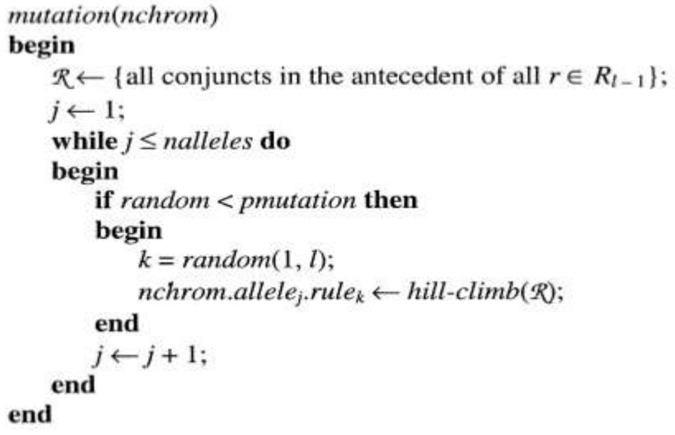
The reproduce function uses a Roulette wheel to select the population [*t*-1] i.e. two different chromosomes, *chrom$_1$* and *chrom$_2$* relative to their value. Then the two chromosomes are passed as parameters to the cross through. Crossover (*chrom* $_1$, *chrom$_2$*) function uses the two-point crossover operator, because it allows the combination of schema. DMEL uses two different strategies for selecting intersections, cross-1and cross-2. Cross-1 operators allow intersections between only two rules, and cross-2 operator allows intersection within only one rule. Fig.3 depicts the genetic operators used by DMEL, implemented in the reproduce function [1].



```
population reproduce(population[$t-1$])
begin
    $chrom_1 \leftarrow$ select(population[$t-1$]);
    $chrom_2 \leftarrow$ select(population[$t-1$]);
    $nchrom_1, nchrom_2 \leftarrow$ crossover($chrom_1, chrom_2$);
    mutation($nchrom_1$);
    mutation($nchrom_2$);
    population $\leftarrow$ steady-state(population[$t-1$], $nchrom_1, nchrom_2$);
    return population;
end
```

Fig.3 Reproduce function

Traditional mutation function is used in reproduce function in DMEL. Fig.4 depicts the detail of mutation function.



```
mutation(nchrom)
begin
    $\mathcal{R} \leftarrow$ {all conjuncts in the antecedent of all $r \in R_{l-1}$};
    $j \leftarrow 1$;
    while $j \leq nalleles$ do
    begin
        if $random < pmutation$ then
        begin
            $k = random(1, l)$;
            $nchrom.allele_j.rule_k \leftarrow$ hill-climb($\mathcal{R}$);
        end
        $j \leftarrow j + 1$;
    end
end
```

Fig.4 Mutation Function

### E. Selection and the Fitness Function

Fitness of chromosome is determined using performance measure which is defined in terms of probability as the value of attribute predicted correctly. Fitness measure is used by the DMEL algorithm to maximize the number of records to be predicted correctly [1].

### F. Criteria for Termination

The following termination conditions:

1) Chromosome performing best and worst in population[*t*] if differs by less than 0.1%, as the population becomes same and there is no improvement in future generation.

2) Terminated on reaching the specified generation by the user.

3) Terminates, current population, no more interesting rules can be identified because it is not likely to find any interesting *lth* order rules, if there is no (*l*-1)th order interesting rules is found [1].

## V. IMPLEMENTATION

In this proposed system HBase acts as a query engine. The entire data of customers is stored in HBase. The scalability of Hadoop is combined by HBase by running on the Hadoop Distributed File System (HDFS) with real time data access as a key-value store and deep analytic capabilities of map reduce [1,2].

HBase plays a key role in reducing the time complexity. The following are the steps of HBase working

1) HBase stores the data on (HDFS) i.e. when user browse the file using file explorer, the

training data is read and it is stored in HBase suitable format

2) A map procedure map the data i.e. values of attributes gets sorted in min max values.
3) Data is shuffled i.e. min value are placed first and then max values of respective attributes
4) Data is reduced using discretization algorithm.

This intermediate output is passed to DMEL algorithm to get the final output.

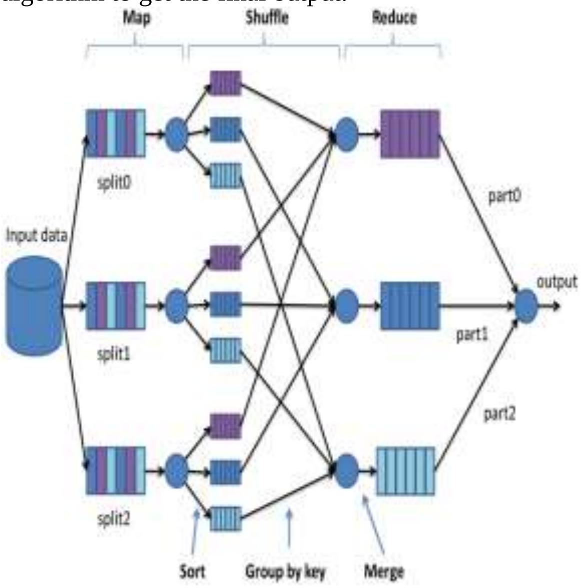


Fig.5 Working of HBase

## VI. Conclusion

There is fierce competition in response to customer retention and loyalty faced by telecommunication industry. In this proposed system we designed a BI tool using DMEL and Hbase approach. DMEL is the algorithm which determines the prediction not only regarding the customers who switch to another subscriber but at the same time the likelihood of the subscriber is also predicted. While predicting bayesian discretization algorithm is used for preprocessing the data. As we all know data in future is going to increase i.e. customers will be increasing so it is difficult to mine the huge data. To overcome this we incorporated DMEL algorithm in Hbase. The huge data is easily stored in Hbase along with garbage values. Hbase mines data considering the garbage value as well so there is no loss of any kind of information during mining. It is expected that time complexity will be reduced using Hbase and at the same time accuracy will be improved to some extent.

## REFERNCES

1) W.H. Au, K.C.C. Chan, X. Yao. A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation 7:6 (2003) 532-545
2) http://a4academics.com/final-year-be-project/11-be-it-cse-computer-science-project/560-data-mining-by-evolutionary-learning-dmel-using-hbase.
3) http://www.informit.com/articles/article.aspx?p=2253412
4) Jonathan L Lustgarten, Shyam Visweswaran, Vanathi Gopal akrishnan and Gregory F Cooper: Application of an efficient Bayesian discretization method to biomedical data: BMC Bioinformatics2011
5) Adnan Idris1, Asifullah Khan, Yeon Soo Lee: Genetic Programming and Adaboosting based churn prediction for Telecom: 2012 IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea
6) http://hbase.apache.org/book/architecture.html
7) http://java.dzone.com/articles/handling-big-data-hbase-part-3
8) http://java.dzone.com/articles/handling-big-data-hbase-part-4
9) http://www.cs.cmu.edu/Groups/AI/html/faqs/ai/genetic/part2/faq-doc-1.html
10) Tom Springer, Charles Kim, Frederic Debruyne, Domenico Azzarello and Jeff Melton: Breaking the back of customer churn, pp 1
11) http://posachaidonut.files.wordpress.com/2013/04/dm_process.gif