# Correlation-based Attribute Selection using Genetic Algorithm

Rajdev Tiwari
ABES Institute of Technology
Campus-II, Vijay Nagar,
Ghaziabad, U.P., (INDIA)

Manu Pratap Singh
Institute of Computer and Information Science
Dr. B.R.Ambedkar University
Khandari, Agra U.P.,(INDIA)

## ABSTRACT

Integration of data sources to build a Data warehouse (DW), refers to the task of developing a common schema as well as data transformation solutions for a number of data sources with related content. The large number and size of modern data sources make the integration process cumbersome. In such cases dimensionality of the data is reduced prior to populating the DWs. Attribute subset selection on the basis of relevance analysis is one way to reduce the dimensionality. Relevance analysis of attribute is done by means of correlation analysis, which detects the attributes (redundant) that do not have significant contribution in the characteristics of whole data of concern. After which the redundant attribute or attribute strongly correlated to some other attribute is disqualified to be the part of DW. Automated tools based on the existing methods for attribute subset selection may not yield optimal set of attributes, which may degrade the performance of DW. Various researchers have used GA, as an optimization tool but most of them use GA to search the optimal technique amongst the available techniques for attribute selection. This paper formulates and validates a method for selecting optimal attribute subset based on correlation using Genetic algorithm (GA), where GA is used as optimal search tool for selecting subset of attributes.

## General Terms:

Data Warehousing, Data Mining, Genetic Algorithms.

## Key words:

Data Warehouse, Attribute subset, Data Source Integration, Correlation Analysis, Relevance Analysis, Genetic Algorithm

## 1. Introduction

Advancements in database and computer techniques have increased the data accumulation speed in many folds. Managing the accumulated data in processed form has become a big challenge for the community. Data Warehouses are aimed to store the preprocessed data to support the Data mining process. Data preprocessing has great impact on the quality of information mined out of the DWs. and has become most essential step of data warehousing for successful data mining [1,2].

### 1.1 Preprocessing of Data

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance [3].

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility [4].

### 1.2 Feature or Attribute selection

Feature selection is a process that selecting a subset from original features. The optimality of a feature subset is measured by an evaluation criterion. As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is usually intractable and many problems related to feature selection have been shown to be NP-hard. A typical feature selection process consists of four basic steps (shown in Figure 1.2.1), namely, subset generation, subset evaluation, stopping criterion, and result validation [4]. In subset generation procedure candidate feature subsets are produced and evaluated. The new one is compared with the previous best one according to a certain evaluation criterion. If the new subset is found to be better, it replaces the previous best subset. This process is repeated until a given stopping criterion is reached [5]. Then the selected best subset is validated on the basis of prior knowledge about the data sets. Feature/attribute selection can be found in many areas of data mining such as classification, clustering, association rules mining, regression etc.

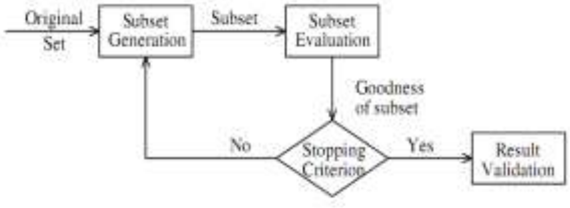

**Figure 1.2.1: Four steps of feature selection**

Feature selection algorithms designed with different evaluation criteria broadly fall into three categories:

Filter model: The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm [6, 7, 8].

Wrapper model: The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model [9, 10].

Hybrid model: The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages. [11,12].

## 1.3 Correlation-based Feature Selection:

Redundancy of attributes in DWs is an important issue. An attribute may be redundant if it can be derived from another attribute or set of attributes. An attribute wish is strongly related to some other attributes are also the redundant ones. Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data. For numerical attributes, we can evaluate the correlation between two attributes, X and Y, by computing the correlation coefficient.

In general, a feature/attribute is good if it is relevant to the class concept but is not redundant to any of the other relevant features. If we adopt the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not highly correlated to any of the other features. In other words, if the correlation between a feature/attribute and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features/attributes does not reach a level so that it can be predicted by any of the other relevant features/attributes, it will be regarded as a good feature/attribute for the classification task. In this sense, the problem of attribute selection requires a suitable measure of correlations between attributes and a sound procedure to select attributes based on this measure. There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory [8]. Under the first approach, the most well known measure is linear correlation coefficient given by the formula:

$$Correlation(r) = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} \qquad 1.3.1$$

Where X and Y are the two features/attributes.

## 1.4 Genetic Algorithm and Attribute subset selection:

The genetic algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. A GA allows a population composed of many individuals (basically the candidates) to evolve under specified selection rules to a state that maximizes the fitness. A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation; crossover combines good strings to generate better offspring; mutation alters a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. The GA cycle continues until the termination criterion is reached.

In feature selection, Genetic Algorithm (GA) is used as a random selection algorithm, Capable of effectively exploring large search spaces, which is usually required in case of attribute selection. For instance; if the original feature set contains N number of features, the total number of competing candidate subsets to be generated is $2^N$, which is a huge number even for medium-sized N. Further, unlike many search algorithms, which perform a local, greedy search, GAs performs a global search [13].

This paper proposes a method for attribute subset selection based of correlation using GA. Correlation between the attributes will decide the fitness of individual to take part in mating. Fitness function for GA is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that are having lower correlation coefficients. In other words we can say lower the correlation is higher the fitness value will be. Rest part of the paper is organized as; section 2 covers a detailed discussion on the related works done so far. Section 3 describes the proposed method with illustrations. Section 4 gives the detail of implementation and result obtained. Section 5 concludes the paper followed by references in section 6.

## 2. Related works:

A number of approaches to feature subset selection have been proposed in the literature, a few of them only are referred here. Most criterion for feature subset selection from the statistics and pattern recognition communities are algorithm independent and do not take into account the difference between the various induction algorithms. The task of finding feature subset that satisfies a given criteria can be describe as state space search. Each state represents a feature subset with the given criteria to evaluate it. Operators determine the partial ordering between the states. In [14] a wrapper based feature selection method is suggested which wraps a number of induction algorithms like holdout, bootstrap and cross-validation for estimating the prediction accuracy. Genetic algorithm approach for feature subset selection appears first in [15]. In 1999, Mark A. Hall advocated about Correlation based feature selection in his Ph.D thesis. According to [16], Inductive learning is one of the major approaches to automatic extraction of useful patterns from massive data. Data become increasingly larger in both number of features (columns) and instances (rows) both in many cases such as genome project text mining, customer relationship management and market basket analysis. This trend poses a severe challenge to inductive learning systems in terms of efficiency and effectiveness. To deal with huge number of instances and large dimensionality, sampling is a common approach. The algorithm suggested by Mark A. Hall (named CFS) exploits heuristic search. Therefore, with quadratic or higher time complexity in terms of dimensionality, existing subset search algorithms do not have strong scalability to deal with high dimensional data. In [8] Lei Yu and Huan Liu came out with new algorithm to overcome the problems of existing algorithms and meet the demand for feature selection for high dimensional data. They developed a novel algorithm based on correlation measure, which can effectively identify both irrelevant and redundant features with less time complexity than subset search algorithms. In 2005 some improved/integrated versions of feature selection algorithms [4, 17] were suggested and validated for their accuracy. In [18], Payam Refaeilzadeh and Lei Tang and Huan Liu have evaluated and compared the performance of Supervised Vector Machines (SVM), Naıve Bayes (NBC), Nearest Neighbor (1NN) and decision tree (C4.5); and three Feature Selection (FS) algorithms: ReliefF (Kononenko 1994), FCBF (Yu & Liu 2003), and information gain (IG).

In late twenties some remarkable efforts have been put in the area of Genetic Algorithm based feature selection approaches. Feng Tan, Xuezheng Fu, Yanqing Zhang,and Anu G. Bourgeois in 2007 [19] proposed a mechnism in which firstly several existing feature selection methods are applied on a data set. Then the feature subsets produced by these methods are fed into the feature pool that is used by the GA in the second stage. Then the GA will try to search an optimal or near optimal feature subset from this feature pool. [20] Presents a new wrapper feature selection method for hyperspectral data, which integrates the Genetic Algorithm and the SVM classifier through properly designed chromosome and fitness function. The purpose is to optimize both the feature subset, i.e. band subset, of hyperspectral data and SVM kernel parameters simultaneously and finally achieve higher classification accuracy. One of the latest methods for GA based attribute selection is briefed in [13]. In this method WEKA [21] GA is used as random search method with four different classifiers namely decision tree (DT) C4.5, Naïve Bayes, Bayes networks and Radial basis function as induction method wrapped with GA.

## 3. Proposed Method:

Dataset for the illustration of proposed method is shown in table 3.1. Proposed GA based algorithm for selection of optimal subset of attributes is diagrammatically represented in figure 3.1.

**Table 3.1: Training Data Tuples**

| RID | Age | Income | Student | Credit Rating | Buys_computer |
|---|---|---|---|---|---|
| 1 | <=30 | High | no | fair | no |
| 2 | <=30 | High | no | excellent | no |
| 3 | 31…40 | High | no | fair | Yes |
| 4 | >40 | Medium | no | fair | Yes |
| 5 | >40 | Low | yes | fair | Yes |
| 6 | >40 | Low | yes | excellent | no |
| 7 | 31…40 | Low | yes | excellent | Yes |
| 8 | <=30 | Medium | no | fair | no |
| 9 | <=30 | Low | yes | fair | Yes |
| 10 | >40 | Medium | yes | fair | Yes |
| 11 | <=30 | Medium | yes | excellent | Yes |
| 12 | 31…40 | Medium | no | excellent | Yes |
| 13 | 31…40 | High | yes | fair | Yes |
| 14 | >40 | Medium | no | excellent | no |

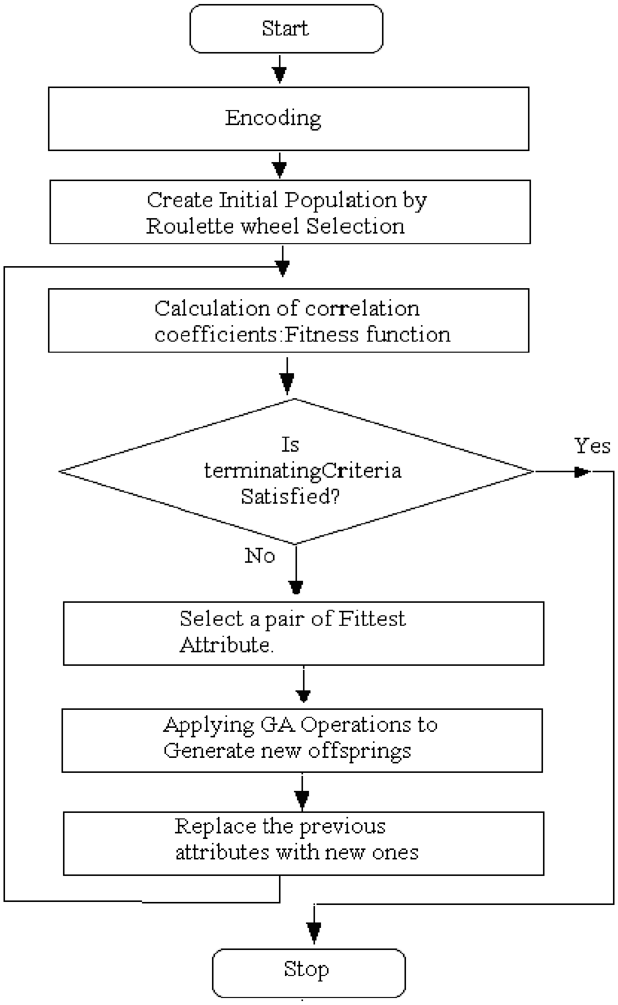
Start
Encoding
Create Initial Population by Roulette wheel Selection
Calculation of correlation coefficients:Fitness function
Is terminatingCriteria Satisfied?
Yes
No
Select a pair of Fittest Attribute.
Applying GA Operations to Generate new offsprings
Replace the previous attributes with new ones
Stop

**Figure 3.1: Proposed method for selecting optimal subset of attributes**

## 3.1 Genetic Representation of attributes (Encoding) and creating initial population:

To apply the GA attributes are required to be encoded as chromosomes so that the GA operations can be applied on them. Encoding is a process of representing individual genes. The process can be performed using bits, numbers, trees, arrays, lists or any other objects. Bit representation of attributes is done over here. Attribute 'Buys_computer' is taken as class label in this particular example and RID is being dropped from the list because it has no role in classification. Rest four attributes namely; Age, Income, Student and Credit Rating are encoded and considered into the initial population, as shown in table 3.1.1.

**Table 3.1.1: Population of Chromosome**

| | **Chromosome Labels** | **Chromosome Strings** |
|---|---|---|
| **Initial Population** | Age | 00 |
| | Income | 01 |
| | Student | 10 |
| | Credit Rating | 11 |

## 3.2 Finding the correlation between attributes and rating them according to their fitness:

Fitness of the individual chromosome is computed on the basis of its correlation coefficients with other attributes in the population. To find out the correlation between the attributes some normalization is done as shown in the table 3.2.1.

**Table 3.2.1: Normalized Dataset**

| RID | Age | Income | Student | Credit Rating | Buys_ computer |
|---|---|---|---|---|---|
| 1 | 1 | 3 | -1 | 1 | -1 |
| 2 | 1 | 3 | -1 | 2 | -1 |
| 3 | 2 | 3 | -1 | 1 | 1 |
| 4 | 3 | 2 | -1 | 1 | 1 |
| 5 | 3 | 1 | 1 | 1 | 1 |
| 6 | 3 | 1 | 1 | 2 | -1 |
| 7 | 2 | 1 | 1 | 2 | 1 |
| 8 | 1 | 2 | -1 | 1 | -1 |
| 9 | 1 | 1 | 1 | 1 | 1 |
| 10 | 3 | 2 | 1 | 1 | 1 |
| 11 | 1 | 2 | 1 | 2 | 1 |
| 12 | 2 | 2 | -1 | 2 | 1 |
| 13 | 2 | 3 | 1 | 1 | 1 |
| 14 | 3 | 1 | -1 | 2 | -1 |

Table 3.2.2 represents the correlation coefficients in the form of matrix along with the minimum coefficient corresponding to every attribute. Minimum is computed on the basis of magnitude only.

As discussed earlier, Fitness function for GA is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that are having lower correlation coefficients. In other words we can say lower the correlation is higher the fitness value will be. The fitness function is taken over here is;

$$f(X) = 1 - \min(r_X) \qquad 3.2.1$$

where, min ($r_x$) is minimum value of correlation coefficient corresponding to any attribute X.

The fitness values of individual chromosomes, computed as discussed earlier, are shown in table 3.2.3.

**Table 3.2.2: Correlation Matrix**

| | Age | Income | Student | Credit Rating | Buys_computer | Min( r ) |
|---|---|---|---|---|---|---|
| Age | 1.00 | -0.42 | 0.17 | 0.00 | 0.18 | 0.00 |
| Income | -0.42 | 1.00 | -0.45 | -0.28 | -0.07 | 0.07 |
| Student | 0.17 | -0.45 | 1.00 | 0.00 | 0.45 | 0.00 |
| Credit Rating | 0.00 | -0.28 | 0.00 | 1.00 | -0.26 | 0.00 |

**Table 3.2.3: Chromosomes with their fitness value**

| Chromosome Label | Chromosome String | Chromosome Fitness | Fitness Ratio(%) |
|---|---|---|---|
| Age | 00 | 1.00 | 25.44 |
| Income | 01 | 0.93 | 23.66 |
| Student | 10 | 1.00 | 25.44 |
| Credit Rating | 11 | 1.00 | 25.44 |

## 3.3 Applying GA Operations and constructing optimal subset of attributes:

Crossover operation with probability, $P_c$=0.8 and Mutation with probability, $P_m$=0.01is applied on the successive populations to produce the new population. The algorithm stops if there is no change to the population's best fitness for a specified number of generations. If the maximum number of generation has been reached before the specified number of generation with no changes has been reached, the process will end. At the end of GA process we have three fittest attributes namely; Age, Student and Credit Rating. This subset of attributes is tested for classification, discussed in the following section.

## 4. Implementation and Result Discussion:

The optimal subset of attribute obtained on executing the proposed algorithm is then tested for its classification accuracy using SIPINA tool of TANAGRA Software. A comparison of the proposed method in terms of classification accuracy is shown in table 4.1. The result shown below is corresponding to Naïve Bayes Classifier for unconditional class distribution.

**Table 4.1: Result Comparison**

| Sno. | Models/Methods | Attribute Subset Selected | % Accuracy |
|---|---|---|---|
| | **Wrapper based Methods** | | |
| 1. | Forward Selection with Cross Validation | (Age, Student) | 80 |
| 2. | Forward Selection with Bootstrap | (Age, Student, Credit Rating) | 90 |
| 3. | Forward selection with Multiple Cross Validation | (Age, Student) | 80 |
| | **Filter based Methods** | | |
| 4. | Relief | (Age, Income Credit Rating) | 90 |
| 5. | Partial Correlation | (Age, Student, Credit Rating) | 90 |
| 6. | Bootstrap Partial Correlation | (Age, Student, Credit Rating) | 90 |
| 7. | MIFS | (Age, Student, Credit Rating) | 90 |
| | **Proposed Method** | | |
| 8. | Correlation-based Attribute Selection using GA | (Age, Student, Credit Rating) | 90 |

## 5. Conclusion:

Results shown previously itself advocate about the capabilities of the proposed method. Specialty of the proposed method can more clearly be stated as follows:

- It performs either equally good or better than many of the existing methods.
- It’s accuracy is more when applied on real and large dataset.
- It is very simple and light because GA is used to search the optimal subset of attributes besides being used for searching the optimal techniques for attribute selections amongst the available ones.

## 6. References:

[1] H. Liu and H.Motoda. Feature Selection for Knowledge Discovery and DataMining. Boston: Kluwer Academic Publishers, 1998.

[2] D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999.

[3] Han & Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2006.

[4] Huan Liu and Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, Pages: 491 - 502 , 2005

[5] H. Liu and H.Motoda. Feature Selection for Knowledge Discovery and DataMining. Boston: Kluwer Academic Publishers, 1998.

[6] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering – a filter solution. In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002.

[7] M.A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 359–366, 2000.

[8] L. Yu and H. Liu. Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, pages 856–863, 2003.

[9] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 247–254, 2000.

[10] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 365–369, 2000.

[11] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 74–81, 2001.

[12] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 601–608, 2001.

[13] M.A.Jayaram , Asha Gowda Karegowda, A.S. Manjunath. Feature Subset Selection Problem using Wrapper Approach in Supervised Learning. International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7, pages 13-16, 2010.

[14] Ron Kohavi and Dan Sommerfield. Feature subset selection using the wrapper method:Overfitting and Dynamic Search Space Technology. First International Conference on Knowledge Discovery and Data Mining, 1995.

[15] Jihoon Yang and Vasant Honavar. Feature suset selection using Genetic Algorithm. IEEE Intelligent Systems, 1998.

[16] Huan Liu, Hiroshi Motoda and Lie Yu. Feature selection with selective sampling. Proceedings of Nineteenth International Conference on Machine Learning, Morgan Kaufman Publishers Inc. pages 395-402, 2002.

[17] Noelia Sanchez-Marono, Amparo Alonso-Betanzos and Enrique Castillo. A New Wrapper Method for Feature Subset Selection. Proceeding of European Symposium on Artificial Neural Networks, Belgium, 2005.

[18] Payam Refaeilzadeh and Lei Tang and Huan Liu. On Comparison of Feature Selection Algorithms. Association for the Advancement of Artificial Intelligence, 2007.

[19] Feng Tan, Xuezheng Fu, Yanqing Zhang,and Anu G. Bourgeois. A genetic algorithm-based method for feature subset selection. Soft Computing - A Fusion of Foundations, Methodologies and Applications, Volume 12 , Issue 2 , Pages: 111 – 120, Springer-Verlag, 2007.

[20] Li Zhuo , Jing Zheng, Fang Wang, Xia Li, Bin Ai and Junping Qian. A Genetic Algorithm based Wrapper Feature selection method for Classification of Hyperspectral Images using Support Vector Machine. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing, 2008.

[21] I. H. Witten, E. Frank.. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufman, San Francisco, 2005.