Knowledge Discovery in Data Mining Using Soft Computing Techniques – A Comparative Analysis

1Rasika Kuware and 2V.P. Mahatme,
3Student (MTech), 2Associate Professor,
1,2Computer Science and Engineering Department, Kavikulguru Institute of Technology & Science, Ramtek, Maharashtra, India.

Abstract: Decades of research have focused on data mining as a multidisciplinary field and also popularly referred to as knowledge discovery from data. In the field of evolutionary computation in data mining provides a balance mixture of theory, algorithm and application. Genetic algorithm is search algorithm based on the mechanics of natural selection, crossover and mutation to reach the optimal solution for better performance. Genetic algorithm is easily parallelizable and has been used for different techniques like as classification, clustering and other optimization problem. Genetic algorithm using clustering is a data stream mining task which is very useful to gain insight of data and its characteristics it comes under the pre-processing step in all mining process.

Clustering is a useful unsupervised data mining task that subdivides an input data set into a desired number of subgroups so that members of the same subgroup will have high similarity and the members of different groups have large differences. K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by its centroid, with the help of to calculate minimum square error function. On the other hand genetic algorithm is one of the commonly used evolutionary algorithm that performs global search to find the optimal solution to a clustering problem. The algorithm typically starts with a set of randomly generated individuals called the population from the dataset and creates successive, new generations of the population by using genetic operations like as natural selection, crossover, and mutation are applied. Each chromosome of the population represents k no. of centroid. The same steps of genetic algorithm are repeatedly applied for a no. of generations to search for appropriate cluster centers in the feature space such that a similarity metric of the resulting clusters is optimized and also do the comparative analysis of the k-means and genetic algorithm, where genetic algorithm shows the better chromosomes or the result as compared to k-means result.

Keywords: Evolutionary Computation, Genetic Algorithm, Clustering, Data Mining, K-Means, Fitness Function.

I. INTRODUCTION

A. Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond respective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies such as Massive data collection, Powerful multiprocessor computers and Data mining algorithms. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Cluster analysis is mainly conducted on computers to deal with very large scale and complex datasets.

The development of computer based techniques, clustering has been widely used in data mining, ranging from web mining, image processing, machine learning, artificial intelligence, pattern recognition, social network analysis, bioinformatics, geography, geology, biology, psychology, sociology, customers behavior analysis, marketing to e-business and other fields. The task of grouping data points into clusters of "similar” items are a form of unsupervised learning that has applications in many fields. For instance, current techniques used for machine vision require processing of digital information obtained from pixels.

The term KDD or Knowledge Discovery in Databases, refers to the broad process of finding knowledge in data, and emphasizes the “high-level” application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.

B. Genetic Algorithm

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a At this juncture, the principal constituents of Soft Computing (SC) are Fuzzy Logic (FL), Neural Computing (NC), Evolutionary Computation (EC) Machine Learning (ML) and Probabilistic Reasoning (PR), with the latter subsuming belief networks, chaos theory and parts of learning theory.

C. Genetic Algorithm and Clustering

The GA aims to use selective ‘breeding’ of the solutions to produce ‘offspring’ better than the parents by combining information from the chromosomes. The GA maintains a population of n chromosomes (solutions) with associated fitness values. Parents are selected to mate, on the basis of their fitness, producing offspring via a reproductive plan. Consequently highly fit solutions are given more opportunities to reproduce, so that offspring inherit characteristics from each parent.

Genetic Algorithms (GAs) are stochastic search methods based on the principle of natural genetic system(Goldberg,1989, Michalewicz, 1992). It performs a multi-dimensional search in order to provide an optimal value of an evaluation (fitness) function in an optimization problem. Unlike conventional search methods, Gas deal with multiple
solutions simultaneously and compute the fitness function values for these solutions. GAs are theoretically and empirically found to provide global near-optimal solutions for various complex optimization problems in the field of operation research, VLSI design, Pattern Recognition, Image Processing, Machine Learning, etc.

While solving an optimization problem using GAs, each solution is usually coded as a binary string (called chromosome) of finite length. Each string or chromosomes is considered as an individual. A collection of P such individuals are called a population. GAs start with a randomly generated population of size P.

In each iteration, a new population of the same size is generated from the current population using two basic operations on the individuals. These operators are Selection and Reproduction. Reproduction consists of crossover and mutation operations.

D. Knowledge Discovery in Databases

Knowledge discovery techniques perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Data mining is an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)  
2. Data integration (where multiple data sources may be combined)  
3. Data selection (where data relevant to the analysis task are retrieved from the database)  
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)  
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)  
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)  
7. Knowledge representation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. A pattern is interesting if

1. It is easily understood by humans,  
2. Valid on new or test data with some degree of certainty,  
3. Potentially useful, and  

The pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge. Several objective measures of pattern interestingness exist. These are based on the structure of the discovered patterns and the statistics underlying them. An objective measure for association rules of the form $X \Rightarrow Y$ is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association.

Although objective measures help identify interesting patterns, they are insufficient unless combined with subjective measures that reflect the needs and interests of a particular user. The main steps in the process of knowledge discovery includes:

1. Business (or Problem) Understanding  
2. Data Understanding  
3. Data Preparation (including all the data cleaning and pre processing)  
4. Modelling (applying machine learning and data mining algorithms)  
5. Evaluation (checking the performance of these algorithms)  
6. Deployment

II. GENETIC ALGORITHM BASED CLUSTERING

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomised, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, especially those follow the principles first laid down by Charles Darwin of "survival of the fittest." Since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones.

It is better than conventional AI in that it is more robust. Unlike older AI systems, do not break easily even if the inputs changed slightly, or in the presence of reasonable noise. Also, in searching a large state-space, multi-modal state-space, or n-dimensional surface, a genetic algorithm may offer significant benefits over more typical search of optimization techniques. (linear programming, heuristic, depth-first, breadth-first, and praxis).

GAs simulates the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consists of a population of character strings that are analogous to the chromosome that see in the DNA. Each individual represents a point in a search space and a possible solution. The individuals in the population are then made to go through a process of evolution.

GAs are based on an analogy with the genetic structure and behaviour of chromosomes within a population of individuals using the following foundations:

- Individuals in a population compete for resources and mates.  
- Those individuals most successful in each ‘competition’ will produce more offspring than those individuals that perform poorly.  
- Genes from good individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent.  
- Thus each successive generation will become more suited to its environment.

A. Search Space

A population of individuals are maintained within search space for a GA, each representing a possible solution to a given problem. Each individual is coded as a finite length vector of components, or variables, in terms of some alphabet,
usually the binary alphabet \{0,1\}. To continue the genetic analogy these individuals are likened to chromosomes and the variables are analogous to genes. Thus a chromosome (solution) is composed of several genes (variables).

A fitness score is assigned to each solution representing the abilities of an individual to ‘compete’. The individual with the optimal (or generally near optimal) fitness score is sought. The GA aims to use selective ‘breeding’ of the solutions to produce ‘offspring’ better than the parents by combining information from the chromosomes. The GA maintains a population of \( n \) chromosomes (solutions) with associated fitness values. Parents are selected to mate, on the basis of their fitness, producing offspring via a reproductive plan.

Consequently highly fit solutions are given more opportunities to reproduce, so that offspring inherit characteristics from each parent. As parents mate and produce offspring, room must be made for the new arrivals since the population is kept at a static size. Individuals in the population die and are replaced by the new solutions, eventually creating a new generation once all mating opportunities in the old population have been exhausted. In this way it is hoped that over successive generations better solutions will thrive while the least fit solutions die out.

New generations of solutions are produced containing, on average, better genes than a typical solution in a previous generation. Each successive generation will contain more good ‘partial solutions’ than previous generations. Eventually, once the population has converged and is not producing offspring noticeably different from others in previous generations, the algorithm itself is said to have converged to a set of solutions to the problem at hand.

**B. Based on Natural Selection**

After an initial population is randomly generated, the algorithm evolves through three operators:

1. **Selection** which equates to survival of the fittest;
2. **Crossover** which represents mating between individuals;
3. **Mutation** which introduces random modifications.

**Cluster wise Euclidean Distance by Genetic Algorithm**

Table 1 output of Genetic Algorithm with \( k=5 \) Cluster wise Euclidean Distance

<table>
<thead>
<tr>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum</td>
<td>TotSum</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7.6364</td>
<td>32.9267</td>
<td>7.6364</td>
</tr>
<tr>
<td>32.7179</td>
<td>65.6446</td>
<td>32.7179</td>
</tr>
<tr>
<td>47.25</td>
<td>112.8946</td>
<td>47.25</td>
</tr>
</tbody>
</table>

**III. K-MEANS BASED CLUSTERING**

The K-means algorithm is one of the partitioning based, non hierarchical clustering methods. Given a set of numeric objects \( X \) and an integer number \( k \), the K-means algorithm searches for a partition of \( X \) into \( k \) clusters that minimizes the within groups sum of squared errors i.e. SSD. The K means algorithm starts by initializing the \( k \) cluster centers. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster.

**A. Steps of K-Means Clustering**

The steps of the K-means algorithm are written below:

1. **Initialization**: Randomly \( K \) data points are chosen to initialize the cluster centers.
2. **Nearest-neighbour search**: Each data point is assigned to the cluster center that is closest to it.

The distance from the data vector to the centroid is calculated using the following equation.

\[
d(z_p, a_j) = \sqrt{\sum_{k=1}^{d} (z_{pk} - a_{jk})^2}
\]

Where,

- \( d \) is the dimension of the data vector,
- \( z_p \) is the centroid of cluster \( p \) and \( a_j \) is the data vector.

3. **Mean update**: New cluster centers are calculated finding the mean of the input vectors assigned to a particular cluster.
4. **Stopping rule**: repeat steps 2 and 3 until no more change in the value of the means

**B. Example of K-Means Clustering**

A data set with 13 constituents found in each of three types of wine with their values present in the wine dataset has been taken for implementing the K-Means algorithm and genetic algorithm. The attributes are

1. Alcohol
2. Malic acid
3. Ash
4. Alkalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Initially with \( k=5 \), five centroid are chosen randomly. Then the Euclidean distances from each data object to the centroid are obtained. For each data object the closest centroid is selected and accordingly the objects are assigned to the clusters.
Cluster wise Euclidean Distance by K-Mean:

The process of comparing the distances from data objects to cluster and updating the centroid value has done for three iterations because after that there is no changes in the new centroid value. The detail implementation of k-means has shown in table-2. The sum of squared distance obtained with this clustering is 112.6996.

Table 2 output of K-Means with k=5 Cluster wise Euclidean Distance

<table>
<thead>
<tr>
<th>Iteration 1</th>
<th>Sum</th>
<th>TotSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25.2903</td>
<td>25.2903</td>
<td>0</td>
</tr>
<tr>
<td>7.6364</td>
<td>32.9267</td>
<td>0</td>
</tr>
<tr>
<td>32.7179</td>
<td>65.6446</td>
<td>0</td>
</tr>
<tr>
<td>47.25</td>
<td>112.8946</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 2</th>
<th>Sum</th>
<th>TotSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45.2903</td>
<td>45.2903</td>
<td>0</td>
</tr>
<tr>
<td>11.2</td>
<td>56.4903</td>
<td>0</td>
</tr>
<tr>
<td>30.9189</td>
<td>87.4093</td>
<td>0</td>
</tr>
<tr>
<td>25.2903</td>
<td>112.6996</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 3</th>
<th>Sum</th>
<th>TotSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45.2903</td>
<td>45.2903</td>
<td>0</td>
</tr>
<tr>
<td>11.2</td>
<td>56.4903</td>
<td>0</td>
</tr>
<tr>
<td>30.9189</td>
<td>87.4093</td>
<td>0</td>
</tr>
<tr>
<td>25.2903</td>
<td>112.6996</td>
<td>0</td>
</tr>
</tbody>
</table>

Performance of Clustering by Genetic Algorithm

The clustering result is given below:

![Figure 1: Performance of Clustering by Genetic Algorithm](image)

Output for Clustering by Genetic Algorithm clusterwise

The cluster wise result is given below:

V. COMPARATIVE ANALYSIS OF K-MEANS AND GENETIC ALGORITHM BASED DATA CLUSTERING

<table>
<thead>
<tr>
<th>K-Means Based Data Clustering</th>
<th>GA Based Data Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: k, dataset, randomly chosen k centroid</td>
<td>Input: k, dataset, P, randomly chosen P chromosomes, tmax</td>
</tr>
<tr>
<td>Objective: Minimizing the sum of squared distance</td>
<td>Objective: Minimizing the sum of distances from each data point to its cluster centroid</td>
</tr>
<tr>
<td>Termination condition: Maximum no. Of iterations reached.</td>
<td>Time complexity: O(tmax<em>p</em>n<em>k</em>d)</td>
</tr>
<tr>
<td>Final clustering may coverage to local optima</td>
<td>GA is based on global search approaches with implicit parallelism</td>
</tr>
<tr>
<td>Time complexity: O(n<em>k</em>d*i) Where n=no. Of data points, k=no.of clusters, d=dimension of data, i=no.of iterations</td>
<td>Time complexity: O(n<em>p</em>n<em>k</em>d) Where n=no.of data points, k=no.of clusters, d=dimension of data, p=population size</td>
</tr>
</tbody>
</table>

CONCLUSION

Clustering is an important unsupervised classification technique where a set of data objects taken in a multi-dimensional space, are grouped into clusters in such a way that data objects in the same cluster are similar in some sense and objects in different clusters are dissimilar in the same sense. K-Means is an intuitively simple and effective clustering technique, but it may get stuck at suboptimal solutions, depending on the choice of the initial cluster centers.

Whereas GA is a randomized search and optimization technique guided by the principles of evolution and natural genetics, and having a large amount of implicit parallelism. So it provides near optimal solutions for objective or fitness function of an optimization problem.

Under limiting conditions, a GA-based clustering technique is expected to provide an optimal clustering, more superior to that of K-Means algorithm, but with little more time complexity.

References


