EMITTER International Journal of Engineering Technology
Vol.2, No.1, June 2014
ISSN: 2443-1168

Hybrid Modeling KMeans – Genetic Algorithms in the Health Care Data

Tessy Badriyah

Electronic Engineering Polytechnic Institute of Surabaya Indonesia
Jl. Raya ITS - Kampus ITS SukoliloSurabaya 60111, INDONESIA
E-mail: tessy@pens.ac.id

Abstract
K-Means is one of the major algorithms widely used in clustering due to its good computational performance. However, K-Means is very sensitive to the initially selected points which randomly selected, and therefore it does not always generate optimum solutions. Genetic algorithm approach can be applied to solve this problem.
In this research we examine the potential of applying hybrid GA-KMeans with focus on the area of health care data. We proposed a new technique using hybrid method combining KMeans Clustering and Genetic Algorithms, called the "Hybrid K-Means Genetic Algorithms" (HKGA). HKGA combines the power of Genetic Algorithms and the efficiency of K-Means Clustering. We compare our results with other conventional algorithms and also with other published research as well. Our results demonstrate that the HKGA achieves very good results and in some cases superior to other methods.

Keywords: Machine Learning, K-Means, Genetic Algorithms, Hybrid KMeans Genetic Algorithm (HGKA).

1. INTRODUCTION

Clustering can be considered the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabeled data. A clustering could be defined as the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data [1].

Among the various clustering algorithms, K-Means (KM) is one of the most popular methods used in data analysis due to its good computational performance. However, it is well known that KM might converge to a local optimum, and its result depends on the initialization process, which randomly generates the initial clustering. In other words, different runs of KM on the same input data might produce different results [2].

In this research, we explored the use of machine learning techniques for health care modeling. A major problem in the health care modeling is a guarantee to get the best model in order to improve the diagnostic with regard to speed, accuracy and reliability.

The main emphasis of this research is combining some techniques in machine learning to improve a classification performance. We propose a hybrid modeling techniques which combines the power of Genetic Algorithms with K-Means Clustering. Our results demonstrate that the Genetic K-Means algorithm is a very competitive and in some cases superior with results from other methods. The performance and efficiency of our approach is investigated using the classification accuracy when compared with conventional method and published results from other researcher.

This paper is organized as follows : in Section 2 we review some related works, in Section 3 we present our proposed, in Section 4 we discuss the experimental results and Section 5 is the conclusions.

2. RELATED WORKS

Many papers have been published on the subject of machine learning and cancer which the majority being concerned with using machine learning methods to identify, classify, detect, other malignancies or disease, especially for diagnosis and prognosis activity. In the following section we summarize selected papers in different research areas of medical data analysis and health care data modeling.

Decision Tree is one of the most well-established classification methods. a very well-known decision tree classifier developed by Ross Quinlan [4]. Many researchers have been used decision tree for classifying medical-related data. Adam et. al. [5] applied decision-tree learning to mass spectra of prostate cancer patients. Another researcher [6] applied an improve decision tree algorithm called T3 for mining stroke related medical data.

Neural Network (NN) are well suited to tackle problems that people are good at solving, like prediction and pattern recognition. Furthermore, NN have been applied within the medical domain for clinical diagnosis [7] and drug development [8].Many clustering algorithms such as K-Means, K-Nearest Neighbour, Fuzzy C Mean and more advance techniques have been successfully applied to various medical task [9]. Joseph Cruz [10] reported several examples of clustering approaches in medical data.

Ensemble methods consist of a set of models and certain model fusion criteria. Statistical, computational and representational reasons have been presented to explain the success of ensemble methods [6]. The fundamental idea is that assuming each classifier's accuracy is better than random. Combining multiple classifiers can improve classification performance if each individual classifier's performance is above an acceptable level (at least better than random) and their outputs are diverse (ideally statistically

independent but in practice different enough such that their errors would not coincide).

3. ORIGINALITY

In this research, we proposed a new technique using hybrid method combining K-Means Clustering and Genetic Algorithms, called the "Hybrid K-Means Genetic Algorithms" (HKGA).

Our new proposed technique HKGA combines the powerful of GA and the simplicity and efficiency of K-Means. Combination between K-Means and GA can be done in two ways. The first method is running GA and then put the results into K-Means. The second method is running K-Means and then use the results as an initial population of GA. Based on published research by [2] the second method is better than the first. By using initial centroid value from GA, at least K-Means clustering start a good first step. We use the second method as the main idea behind our proposed technique for combine K-Means Clustering and GA.The uniqueness of this research can be identified by using GA as initialization centroid value to optimize K-Means clustering.

4. SYSTEM DESIGN

First, we are going to overview two existing methods: K-Means Clustering and Genetic Algorithms, and then give a detail explanation about proposed techniques: a Hybrid KMeans Genetic Algorithms (HKGA).

4.1 K-Means Clustering

K-Means algorithm is relatively faster, simpler and needs less computation, but it has some weaknesses. The disadvantages K-Means algorithm is that it is sensitive to the initially selected points, and therefore it does not always generate the same results. This algorithm also does not guarantee to find the global optimum.

K-Means Algorithms can be explained in detail in the following algorithm [Algorithm 1.].

Algorithm 1. K-Means Clustering
Input:
k: the number of clusters,
D: a data set containing n objects.
Output: A set of k clusters.
Method:
(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4)update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

4.2 The Genetic Algorithms

The Genetic Algorithm (GA) technique was originally proposed by Holland [12]. GA has been applied to many function optimization problems and is shown to be good in finding optimal or near optimal solutions. The GA can be explained in detail in the following algorithm (Algorithm 2)

Algorithm 2. Genetic Algorithms
```
begin
    t = 0;
    initialize P(t);
    evaluate structures in P(t);
    while termination condition not satisfied do
    begin
        t = t + 1;
        select_repro C(t) from P(t-1);
        recombine and mutate structures in C(t) to C'(t);
        evaluate structures in C'(t);
        select_replace P(t) from C'(t) and P(t-1);
    end
end
```

4.3 Framework of Proposed Technique

The originality of this research has been explained in section 3, and the following we will give a detail explanation about our proposed techniques: a Hybrid modelling K-Means Genetic Algorithms (HKGA).

The Hybrid K-Means Genetic Algorithms is explained in the following detail (Algorithm 3)

Algorithm 3. Framework of Hybrid K-Means Genetic Algorithms
```
Stage 1: Generate Initial Population from K-Means
Input:
D: a data set
S: size of population (number of chromosome)

Output: A set of N chromosome.
Method:
(1) Determine number of chromosome (population size) as S
(2) Repeat
(3) Randomly choose initial cluster center
(4) Running K-Means Algorithm on Dataset D
(5) until S;

Stage 2: Optimize using Genetic Algorithm
Input:
D: a data set
Genetic Parameter: crossoverrate, mutationrate, size of population
maxIter : maximum iteration
```

Algorithm 3. Framework of Hybrid KMeans Genetic Algorithms (continued)
……

Output: The best chromosome.
Method :
(1) repeat
(2) parent selection using roulette wheel based on fitness function
(3) crossover process;
(4) mutation process
(5) survivor selection, pass the selected chromosome to the next generation\
(6) until maxIter;

We will describe in detail, the elements of GA in our proposed techniques.

4.3.1 Chromosome Representation

In our approach, each chromosome represents sequence number of centroid representation. For example, the problem which have two classes (benign or malignant), the chromosome consists of 2 centroids (number of clusters or classes). The length of chromosomes indicates the number of gen in the chromosomes. For example, the dataset consist of 9 attribute and have 2 classes, we can calculate the length of the chromosome is = 9 x 2 = 18 genes.

Cluster-1	Attribute 1	Attribute 2	Attribute 3		Attribute-n
Cluster-2	Attribute 1	Attribute 2	Attribute 3		Attribute-n
	.....	.....	.....		.....
Cluster-k	Attribute 1	Attribute 2	Attribute 3		Attribute-n

k = number of clusters
n = number of attributes

Figure 1. Structures of the chromosomes for HKGA

4.3.2 Initial Population

In the initialization phase, initial population of N chromosomes is generated. N is the size of population means the number of solution (number of chromosome) in population. To generate chromosomes (c1; c2; . . . ; cn), we employ the K-Means Clustering.

4.3.3 Fitness value calculation

The following is the pseudo code to calculate fitness value In this pseudo code, we need total fitness value when use roulette wheel selection to choose which parent involved in crossover or mutation process.

Algorithm 4. Fitness value calculation

```
total_fitness=0;
for i=1: number_of_chrom,,
  [1] get accuracy between output and datatarget
  [2] set accuracy as fitness value
  [3] total fitness value is sum up of all fitness value
end;
```

4.3.4 Selection

We use Roulette Wheel Selection in the Selection process, which the pseudo code in explained below:

Algorithm 5. Roulette Wheel Selection

```
function individu=Selection(number_of_chrom,kromosom,V)
[1] total_fitness=0;
[2] Accumulate fitness value from all the chromosome
        Save to total_fitness
[3] for i=1:number_of_chrom,
        Calculate the proportion each of fitness value to  total_fitness
end;
[4] for j=1:number_of_chrom,
      Generate random number and see the position based on proportion of
       fitness value of each chromosometaken in accordance with the individual
       chromosome selected
end;
```

4.3.5 Crossover

We use 2-point crossover to generate 2 offspring. (2 point obtained at random).

4.3.6 Mutation

For mutation operator we use changes in the small value on gen values in the chromosomes. The mutation process is described as follows:

Algorithm 6. Mutation Process
```
[1] for i=1:number_of_chrom,
      Let (a1,a2, …, ank) denote gen of the chromosome;
[2]   for i=i:length_of_chrom,
[3]       v = generate random number between [0,1]
[4]       if (v<=mutation_rate) then do
[5]          Generate random number between [0,1]
[6]          For the i-th point of the chromosome;
              Increment/Decrement aj with the smallest
              number (0.1)
      End;
    End;
  End;
```

4.3.7Survivor Selection

We believe that diversity plays an important role in the Genetic Algorithms, thus maintaining diversity in the population is a point we can conclude. In the following, pseudo code in Selection process:

Algorithm 7. Selection process
Pseudo code for Survivor Selection:
[1] Listing the parent chromosome and pick the best one.
[2] Sorting chromosome combine parent and child (offspring)
[3] Pick the remain (N-1) from the top after sorting

5. EXPERIMENT AND ANALYSIS

5.1 Dataset

In this section, we describe dataset used to evaluate the proposed method. We use the following dataset:

(1) Breast Cancer from MATLAB
There are 9 attributes used to determine whether a patient classified into benign or malignant (2 classes), and there are 699 cases in the dataset.

(2) Thyroid dataset from MATLAB
This data set has 3 classes, contains information related to thyroid dysfunction. The problem is to determine whether a patient has a normally functioning thyroid, under functioning (hypothyroid), or over-active functioning hyperthyroid. There are 7200 cases in the data set, with 21 attributes used to determine to which of the three classes the patient belongs.

(3) Lung Cancer
The dataset is taken from UCI Machine Learning Repository [15]. There are 56 attributes and 2 classes with 32 instances.

(4) Wisconsin Breast Cancer (WBC) Dataset
The breast cancer data set is also available in UCI Lab [15] and it was obtained from the University of Wisconsin Hospitals, Madison from Dr.

William H. Wolberg. This dataset is widely used among researchers to test the effectiveness of classification algorithms. The aim of the classification is to distinguish between benign and malignant cancers (2 classes) based on 9 attributes, and there are 683 instances.

5.2 Experiment Procedure

For the experiment, we use cross validation methods. In cross-validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The accuracy of the system is the average accuracy over the 10 iterations [6]. We execute every algorithms on four different dataset ten times and taken the average value and standard deviation.

5.3 Experimental Results

The following tables show the comparison results between the conventional algorithms and the hybrid algorithm. Table 1 shows the results of experiments with K-Means Algorithm.

Table 1. K-Means algorithms results

No	Dataset Name	Best Result	Worse Result	Average ± standard deviation
1	Breast Cancer	95.85	95.70	95.74±0.06
2	Thyroid	86.51	51.90	72.35±10.65
3	Lung Cancer	77.78	51.85	65.93 ± 8.87
4	WBC Dataset	98.98	96.49	96.74 ± 0.79

We used the same input parameters for HKGA: number of chromosomes=10, probability of crossover= 0.95, probability of mutation = 0.7 and maximum iteration = 50. There is no reason to set parameters like that. We just set it up randomly.

Table 2 shows the experiment by using genetic algorithms.

Table 2. Experiment resultsusing genetic algorithm

No	Dataset Name	Best Result	Worse Result	Average ± standard deviation
1	Breast Cancer	96.99	92.56	95.11 ± 1.52
2	Thyroid	92.61	92.53	92.58±0.02
3	Lung Cancer	92.59	74.07	83.33 ± 5.59
4	WBC Dataset	99.85	96.78	99.27 ± 0.93

Table 3 uses a hybrid scenario 1 in which the genetic algorithm is run first and then KMeans, called by HGAK (Hybrid Genetic Algorithms - KMeans)

Table 3. Experiment results using Hybrid Genetic K-Means (HGAK)

No	Dataset Name	Best Result	Worse Result	Average ± standard deviation
1	Breast Cancer	95.85	95.71	95.84 ±0.05
2	Thyroid	92.61	92.53	92.58±0.02
3	Lung Cancer	88.89	74.07	84.81 ± 4.43
4	WBC Dataset	99.12	98.98	99.06±0.076

Table 4 uses a hybrid scenario 2 where the K-Means algorithm run first and then the results are used as an initial population for the next run of Genetic Algorithm.

Table 4. Experiment resultsusing Hybrid K-Means Genetic Algorithm (HKGA)

No	Dataset Name	Best Result	Worse Result	Average ± standard deviation
1	Breast Cancer	97.14	95.85	96.22 ± 0.50
2	Thyroid	99.61	94.36	96.69 ± 1.77
3	Lung Cancer	96.3	92.59	93.70 ± 1.79
4	WBC Dataset	99.71	99.27	99.46 ± 0.22

Table 5 shows that our proposed method Hybrid K-Means Genetic Algorithms (HKGA) achieved the best result compare to other conventional algorithms. These results show that the combination of K-Means algorithms and Genetic Algorithms give better results than each algorithm run independently. Only in one case (in WBC dataset), Genetic Algorithms performs better then Hybrid Genetic Algorithms K-Means (HGAK). In Table 5 we also can see that the accuracy between GA and HGAK is the same in the thyroid dataset.

Table 5 : Comparison between conventional methods and hybrid method

No	Dataset Name	K-Means	GA	HGAK	HKGA
1	Breast Cancer	95.74±0.06	95.11±1.52	95.84±0.05	96.22±0.50
2	Thyroid	72.35±10.65	92.58±0.02	92.58±0.02	96.69±1.77
3	Lung Cancer	65.93±8.87	83.33±5.59	84.81±4.43	93.70±1.79
4	WBC Dataset	96.74±0.79	99.27±0.93	99.06±0.076	99.46±0.22

Using the same dataset, we compare our proposed algorithm Hybrid K-Means Genetic Algorithm (HKGA) with other published research. All of the algorithms given in Table 6 used the same dataset (The Wisconsin Breast Cancer Dataset).

Table 6. Classification accuracy benchmark forWisconsin Breast Cancer Dataset

No	Researchers	Methods	Accuracy
1	Biying Zhang [16]	Novel Neural Network	97.24%
2	A. Verikas, et.al [17]	fuzzy derivative	97.10%±0.75
3	Akay, MF. [18]	F-score + SVM	99.51%
4	Our Result	HKGA	99.46%±0.22

Table 6 shows that our proposed method (HKGA) has a better accuracy (99.46%) than Novel Neural Network method (97.24%) proposed by [16] and Fuzzy Derivative method (97.1) proposed by [17]. Unfortunately our method is still below F-score & SVM method (99.51) proposed by [18].

6. CONCLUSION

In this research we have presented two scenario of a new machine learning approach using hybrid method combining K-Means Clustering and Genetic Algorithms, called the the "Hybrid Genetic Algorithms – K-Means (HGAK)" and "Hybrid K-Means - Genetic Algorithms" (HKGA).

The proposed method can be characterized by the design of its operators, including encoding, crossover, mutation and selection survivor. Among the two hybrid modeling scenarios proposed in this study, HKGA shows a better performance than HGAK. In HKGA, we treated the KMeans Clustering to generate initial population in Genetic Algorithm and used GA to solve the problem in order to obtain global optimal solution.

Hybrid Genetic Algorithms K-Means (HGAK) does not always improve the result of conventional algorithms, it can produce similar result or even worse. It's depends on the characteristics of the data itself.

Similar to K-Means, GA are also having problems trapped in premature convergence. To avoid the convergence process of the algorithm, we used survivor selection strategy to keep diversity in the population of GA. The element of GA that proposed in this research has worked well.

Based on experimental results, we can conclude that HKGA that proposed in this research provides a satisfactory performance compared with the other algorithm (K-Means, Genetic Algorithms, and HGAK), and outperform compared with some algorithms from the other researchers.

There are two important issues in GA: exploitation and exploration. When we get the better offspring which have better fitness value that means exploitation, sometime some of the good chromosome dominates the population, than GA will trapped to premature convergence. So the most important thing is to keep the diversity in the process of survivor selection to give the population of GA have a chance to do an exploration to find a global optimum.

In the future works, we have a plan to combine GA with other techniques in Machine Learning that never done before. More investigations should be done and considered to carry out a comprehensive and comparative study on hybrid modelling techniques in the health care data.

REFERENCES

[1] Boncheva, V.M., **Using the Agglomerative Method of Hierarchical as A Data Mining Tool in Capital Market**, *International Journal Information Theories & Applications*, Vol.15, 2008.

[2] Al-Shboul, B., Myaeng, S.H., **Initializing K-Means Using Genetic Algorithms, World Academy of Science**, *Engineering and Technology,* Vol. 54, 2009.

[3] Delen, D. W.,**Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods**,*Artificial Intelligence in Medicine*, 2004.

[4] Quinlan, J.,**Improved use of Continuous Attributes in C4.5**,*AI Research*, pp. 77-90, 1996.

[5] Adam B-L, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH et al.. **Serum Protein Finger Printing Coupled with A Pattern-Matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men**,*Cancer Research*, 2002.

[6] Tjortjis,C., Saraee,M., Theodoulidis,B., Keane, J.A.,. **Using T3, an Improved Decision Tree Classifier for Mining Stroke Related Medical Data**, University of Manchester, 2005.

[7] Guan W, Zhou M., Hampton C.Y., Benigno B.,**Ovarian Cancer Detection from Metabolomic Liquid Chromatography/Mass Spectrometry Data by Support Vector Machine**, BMC Bioinformatics, 2000.

[8] Weinstein J, K. Kohn and M. Grever, et. al.. **Neural Computing in Cancer Drug Development : Predicting Mechanism of Action**,*Information Science* , pp. 447-451, 1992.

[9] Dhiraj,K., Rath,S.K.,**Gene Expression Analysis Using Clustering**. *International Journal of Computer and Electrical Engineering*, 2009.

[10] Cruz, J.A., David S. Wishart,**Applications of Machine Learning in Cancer Prediction and Prognosis**, *Journal on Cancer Informatics*, pp. 59-78, 2006.

[11] Jiawei Han, M. K.,**Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, 2000.

[12] Holland, J. H., **Adaptation in Natural and Artificial Systems**, AnnArbor, MI: Univ. of Michigan Press, 1975.

[13] De Jong, K.A., **Evolutionary Computation: A Unified Approach**, MIT Press, Cambridge, MA, 2006.

[14] Chen, H.; Fuller, S.S.; Friedman, C.; Hersh, W.,**Knowledge Management and Data Mining in Biomedicine**,*Medical Informatics*, 2005.

[15] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/machine-learning-databases/

[16] Zhang, B,. **A Joint Evolutionary Method Based on Neural Network for Feature Selection**, *Second International Conference on Intelligent Computation Technology and Automation(IEEE)*, 2009.

[17] A. Verikasa,M. Bacauskiene, D. Valincius, A. Gelzinis,**Predictor Output Sensitivity and Feature Similarity-Based Feature Selection**,*Science Direct : Fuzzy Sets and Systems*, 2008 .

[18] Akay, M. F.,**Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis**,*Science Direct: Expert Systems with Applications*, 2009