A methodology to find clusters in the data based on Shannon's Entropy and Genetic Algorithms

Edwin Aldana-Bobadilla[1], **Angel Kuri-Morales**[2]
[1]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, Mexico City, Mexico
[2]Departamento Académico de Computación, Instituto Tecnológico Autónomo de México, Mexico City, Mexico

Abstract - *The most common clustering methods are based on metrics that allow the determination of the similarity between elements of a given data set. This similarity allows us to divide the data set into subsets (clusters) that contain "highly similar" elements. The use of a metric imposes two constraints. First, the shape of the found clusters is generally hyper-spherical (in the space of the metric) due to the fact that each element in a cluster lies within a radial distance relative to a given center. Second, the metric may be sensitive to the probability density function of the data set. Following this fact several methods based on statistical approaches have become an attractive and powerful option. These involve the estimation of the probability density function (pdf) of the data set which minimizes an optimality criterion. Generally this is a highly non-linear and usually non-convex optimization problem which disallows the use of traditional optimization techniques. In this paper we propose a statistical method based on Shannon's Conditional Entropy which uses a rugged genetic algorithm to find the optimal pdf. Each individual of the Genetic Algorithm is a possible solution of a clustering problem. The fitness of an individual is determined by Shannon´s entropy encoded in its genome and an additional constraint related to the "quality" of this solution. The "quality" is measured through a validity index of the clustering process. A novel and important aspect of our method is the form of representation of the objects of the data set in order to reduce the computational complexity due to the high dimensionality. We show that our proposal has high effectiveness relative to methods as k-means, fuzzy c-means and Kohonen Maps with a synthetic data set.*

Keywords: Clustering, Information Theory, Genetic Algorithms, Bayesian Classifier, Data Mining.

1 Introduction

The clustering process is an optimization problem that maximizes the similarity between objects or elements that belong to same cluster and minimizes the similarity between elements of different clusters. The effectiveness of a clustering method is given by several factors such as the metric and the desired number of clusters.

Particularly, the use of a metric imposes some constraints on the shape of clusters found. These shapes generally are hyperspherical (in the space of the metric) due to the fact that each element in a cluster lies within a radial distance relative to a given center. In other words the elements of a cluster tend to group around a single mean value (center) which sometimes disallows the extraction of hidden patterns in the data set.

In this paper we propose an alternative method based on a statistical approach. Our proposal does not use explicitly a metric to determine the elements that belong to given cluster. Overall, this proposal is an iterative search of a partition model of the data set in which the entropy (uncertainty) is minimized. In order to determine the entropy of the data set for a particular partition model, the estimation of its probability density function (pdf) is necessary. This estimation can be achieved statistically from three different methods: parametric, semi-parametric and non-parametric [15]. Unlike parametric and semi-parametric methods, the non-parametric methods do not make any assumption about of the pdf of the data set. The Parzen window [5] is among the most widely-used non-parametric density estimation method.

Different clustering methods have been proposed around these non-parametric methods and minimum entropy principle [9], [15],[16]. These methods can be seen as an iterative search of an optimal pdf of the data set such that the entropy is minimal. However, depending on the dataset the search may be unfeasible or may yield local optimal solutions. Thus, this is a highly non-linear and usually non-convex optimization problem which disallows the use of traditional optimization techniques or pdf estimation methods.

We propose a method which uses a rugged genetic algorithm (the so-called Vasconcelos's GA [12]). Each individual of GA is a possible solution of a clustering problem which represents a pdf of the data set. The fitness of an individual is based on the minimum entropy principle and an additional constraint related to the "quality" of the solution. The "quality" is measured through an validity indices of the clustering process. Several validity indices have been developed and introduced [4],[8], [11]. A novel and important aspect of our proposal is the form of representation of the objects of the data set. Generally the properties of each object are represented as real values of vector in a Euclidean space. The dimensionality of this vector is given by the number of such properties. Its value is an important element of the computational complexity of a clustering algorithm.

In order to reduce the dimensionality, statistical techniques as such as Pearson's correlation analysis [3] and/or principal components' analysis [17] have been used. In many cases, however, these techniques are sensitive to the data distribution and impair the effectiveness of the clustering process. To avoid this fact we map the *n*-dimensional vector space of the data set to the space of all possible strings (words) that can be built using the symbols of an alphabet ∑. This transformation allows us to represent an object of data set as a word of length *n* (for a *n*-dimensional space) and a cluster as a subset of words with some "degree of similarity". The entropy of a cluster is determined by the probability distribution of all words that belong to that cluster.

Our work begins with an account of several concepts which are needed to expose our method. Then, we expound the fundamental process of our proposal. Finally we show several numerical results and the respective conclusions.

2 Theoretical Aspects

In what follows we make a very brief mention of the theoretical aspects having to do with the proper understanding of our proposal. The reader may find more details in the references.

2.1 Minimum Entropy Principle

Shannon's entropy [20] allows us to measure the uncertainty associated with a random variable *X*. Mathematically, Shannon's entropy of *X* with a probability mass function *p(x)* is defined as:

$$H(X)=-\sum p(x)\log(p(x)) \qquad (1)$$

The possible values of a random variable *X* occur with certain probability *p(X=x)* or simply *p(x)*. When *p(x)* is uniformly distributed we say that the uncertainty is greatest or that the process represented by the random variable *X* has a highest degree of "disorder". Figure 1 represents the entropy for two possible values of *X* with probabilities *p* and *1- p*; when *p=0.5* the entropy is maximum.

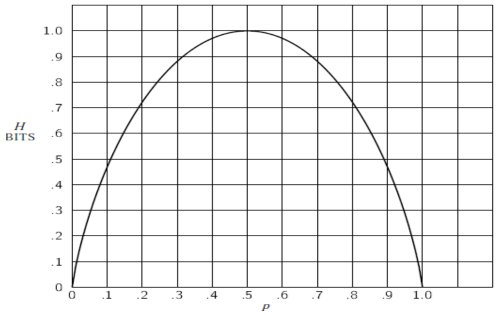


Figure 1. Entropy in the case of two possibles values with probabilities *p* and *(1-p)*

In the context of the clustering problem we assume that a cluster is a subset of the data set which has minimum entropy. It means that a cluster is a partition of data set with minimum degree of "disorder". The entropy of a cluster is directly related to its elements. In terms of probability, the entropy of the cluster depends of the pdf of its elements. In what follows we expound on this fact.

Let *D* be the data set with *K* partitions (clusters) and *x* an element that belong to *D*. Then the pdf of *x* is given by:

$$p(x)=\sum_{i}^{K} p(x|i)\,p(i) \qquad (2)$$

where *p(i)* is the prior probability for the *i-th* partition and *p(x|i)* is the prior probability of *x* given the *i-th* partition. In Figure 2 we show an intuitive representation of the probabilities *p(x|1)* and *p(x|3)*, the probability *p(x|2)* is zero.

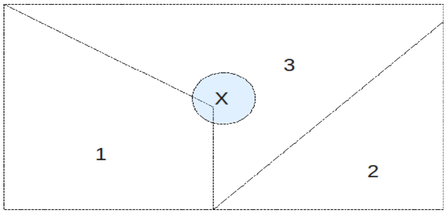


Figure 2. Probability space of a data set with three partitions. The element *x* belongs to partition 1 and 2 with a probability greater than zero.

However we would like to know the dependence of pdf of the *i-th* partition with respect to *x*. This dependence is given by Bayes Theorem [10] :

$$p(i|x)=\frac{p(x|i)\,p(i)}{p(x)} \qquad (3)$$

When *p(i|x)* is uniformly distributed for all i, we can say that the element *x* belongs to any partition and thus the uncertainty is maximum (see Figure 3a.). On the other hand if all *p(i|x)* but one are zero (one having the value unity) then we are certain of the partition to which *x* belongs (see Figure 3b).

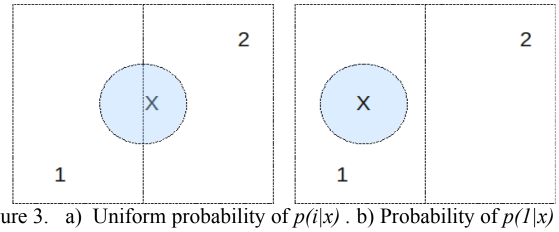


Figure 3. a) Uniform probability of *p(i|x)* . b) Probability of *p(1|x)*

Now, let *C* be a random variable whose possible values are *1,2,..K* which represent the partitions of *D*. Let *X* be a random variable whose possible values are all elements *x* that belong to *D*. Then the entropy of *C* given *X* is:

$$H(C|X)=-\sum_{i=1}^{K} p(i|x)\log(p(i|x)) \qquad (4)$$

where *p(i|x)* is a posteriori pdf. Thus, our goal is to find this function such that *H(C|X)* is minimum. For reasons already mentioned we use a genetic algorithm. The entropy given by Equation 4 is called *Conditional Entropy* [1].

2.2 Genetic Algorithms

Genetic Algorithms (an interesting introduction to GA's and other evolutionary algorithms may be found in [2]) are optimization algorithms which are frequently cited as "partially simulating the process of natural evolution". Although this a suggestive analogy behind which, indeed, lies the original motivation for their inception, it is better to understand them as a kind of algorithms which take advantage of the implicit (indeed, unavoidable) granularity of the search space which is induced by the use of the finite binary representation in a digital computer. In such finite space, numbers originally conceived as existing in R^n actually map into B^m space. Thereafter it is simple to establish that a genetic algorithmic process is a finite Markov chain (MC) whose states are the populations arising from the so called genetic operators: (typically) selection, crossover and mutation [19]. As such they display all of the properties of a MC. From this fact one may prove that:

1. The final results of the evolutionary process are independent of the initial population and

2. A GA preserving the best individual arising during the process will converge to the global optimum (albeit the convergence process is not bounded in time).

Their most outstanding feature is that, as opposed to other more traditional optimization techniques, the GA iterates simultaneously over several possible solutions. Then, other plausible solutions are obtained by combining (crossing over) the codes of these solutions to obtain hopefully better ones. The solution space (SS) is, therefore, traversed stochastically searching for increasingly better plausible solutions. In order to guarantee that the SS will be globally explored some bits of the encoded solution are randomly selected and changed (a process called mutation). The main concern of GA-practitioners (given the fact that well designed GAs, in general, will find the best solution) is to make the convergence as efficient as possible. The work of Forrest et al. has determined the characteristics of the so-called Idealized GA (IGA) which is impervious to GA-hard problems [6].

2.2.1 Vasconcelos's Genetic Algorithm

The implementation of the IGA is unattainable in practice. However, a practical approximation called the Vasconcelos's GA (VGA) has been repeatedly tested and proven to be highly efficient [12]. The VGA, therefore, turns out to be an optimization algorithm of broad scope of application and demonstrably high efficiency. A statistical analysis was done by minimizing a large number of functions and comparing the relative performance of six optimization methods of which five are GAs[1]. The ratio of every GAs absolute minimum (with probability $p = 0.95$) relative to the best GAs absolute minimum may be found in Table 1 under the column "Relative Performance". The number of functions which were minimized to guarantee the mentioned confidence level is shown under "Number of Optimized Functions". It may be seen that VGA, in this study, was the best of all the analyzed variations. Interestingly the CGA (the classical or "canonical" genetic algorithm) comes at the bottom of the list with the exception of the random mutation hill climber (RHC) which is not an evolutionary algorithm. According to these results, the minimal found with VGA are, in the worst case, more than 25% better than those found with the CGA. Due to its tested efficiency, we now describe in more detail VGA.

As opposed to the CGA, VGA selects the candidate individuals deterministically picking the two extreme (ordered according to their respective fitness) performers of the generation for crossover. This would seem to fragrantly violate the survival-of-the-fittest strategy behind evolutionary processes since the genes of the more apt individuals are mixed with those of the least apt ones. However, VGA also retains the best *n* individuals out of the *2n* previous ones.

Table 1: Relative Performance of Different Breeds of Genetic Algorithms

Algorithm	**Relative Performance**	**Number of Optimized Functions**
VGA	1.000	2,736
EGA	1.039	2,484
TGA	1.233	2,628
SGA	1.236	2,772
CGA	1.267	3,132
RHC	3.830	3,600

The net effect of this dual strategy is to give variety to the genetic pool (the lack of which is a cause for slow convergence) while still retaining a high degree of elitism. This sort of elitism, of course, guarantees that the best solutions are not lost. On the other hand, the admixture of apparently counterpointed plausible solutions is aimed at avoiding the proliferation of similar genes in the pool. In nature as well as in GAs variety is needed in order to ensure the efficient exploration of the space of solutions. As stated before, all elitist GAs will eventually converge to a global optimum. The VGA does so in less generations. Alternatively we may say that VGA will outperform other GAs given the same number of generations. Besides, it is easier to program because we need not simulate a probabilistic process. Finally, VGA is impervious to negative fitness's values. We, thus, have a tool which allows us to identify the best values for a set of predefined metrics possibly reflecting complementary goals. For these reasons we use in our work VGA as the optimization method. In what follows we explain our proposal based in the concepts mentioned above.

[1] VGA: Vasconcelos' GA; EGA: Eclectic GA; TGA: Elitist GA; SGA: Statistical GA; CGA: Canonical (or Simple) GA; RMH: Random Mutation Hill Climber.

3 Methodology

We begin our explanation by discussing the preprocessing of the data set. It will allow us to change the vector representation of the data in order to facilitate subsequent calculations. Second, we show the details of the genome's encoding in the context of the clustering problem. Finally we show the way to evaluate each solution or individual in order to find the best.

3.1 Preprocessing of the data set.

Let Σ be an alphabet and w a string that contains symbol of Σ. Let D be a data set. Let $x_i = \{a_1, a_2, ... a_n\}$ be an n-dimensional vector such that $x_i \in D$ where $a_i \in \boldsymbol{R}$ and $D \in \boldsymbol{R}^n$.

Let $\perp a_k$, $\top a_m$ be the minimal and maximal value $\forall\ a_i \in D$. Let Δ be the difference between $\top a_m$ and $\perp a_k$, then we assign to every symbol of Σ an interval value as following:

Table 2: Assigning values to symbols of Σ

Symbol	Interval Value
s_o	$\left[.\perp a_k, \perp a_k + \frac{\Delta}{\lvert\Sigma\rvert} \right]$
s_1	$\left[s_{0\,max}, s_{0max} + \frac{\Delta}{\lvert\Sigma\rvert} \right]$
...	...
s_m	$\left[s_{m-1\ max}, s_{m-1\ max} + \frac{\Delta}{\lvert\Sigma\rvert} \right]$

Where $s_{i\ max}$ is the maximum interval value of S_i and $|\Sigma|$ is the cardinality of Σ (m=$|\Sigma|$). Now we assume that Σ is conformed by the letters of the English alphabet and $\top a_m$=1 and $\perp a_k$=0. In accordance with Table 2 we can determine the interval values of Σ as shown in Table 3.

Table 3: Possible assignment of values for letters of the English alphabet

Symbol	Symbol Value
A	$\left[0, 0 + \frac{1}{26} \right]$
B	$\left[\frac{1}{26}, \frac{1}{26} + \frac{1}{26} \right]$
...	...
Z	$\left[\frac{25}{26}, 1 \right]$

Moreover, if we assume a data set D in R^3 such that some x=[0.038,0.022,0.99]. Then x may be represented by w=AAZ. Thus, $\forall x \in D$, $\exists w \in \Sigma^*$. We represent the set of all strings or words w as D'. For practical reasons we use the English Alphabet although the method described does not depend on any particular symbol set. However this method will be affected by the cardinality of Σ. For example, if $|\Sigma|$ $=1$ we have that all elements of the data set are represented by the same word regardless of their degree of similarity. Otherwise when the value of $|\Sigma|$ is higher we will have more precision but the performance will be affected.

3.2 Encoding of the genome.

The individuals of the algorithm have been encoded as follows. a) The length of the genome is equal to the cardinality of D'. b) Each gene is associated with a word of D'. The value of each gene corresponds to a label (for practical purposes we use $1,2,...K$) of the cluster to which the word belongs. Thus, the i-th gene represent the cluster to which the i-th word belongs. Figure 4 exemplifies a genome for K=3.

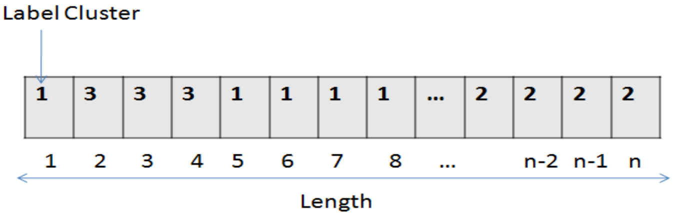


Figure 4. Genome of the individual (K=3)

3.3 Fitness

Each individual is a possible solution of a clustering problem which is evaluated through a fitness function. In what follows we explain how this function is defined in the context of our method.

Suppose that D'=$\{AAA, ACA, MOM, NPM, ADE, UVT, VXT, NQP, VWV\}$ and K=3. Let I_i be the i-th individual of the population whose genome are shown in Figure 5.

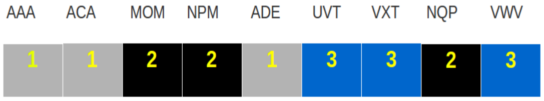


Figure 5. Possible solution given by an Individual for K=3. Here are shown the words associated to each gene.

As discussed above we use the Minimum Entropy Principle. In Equation 4 X is a random variable whose set of possibles values belongs to D. Thus, if the data set D is transformed to set D' (conformed by words w) then Equation 4 may be rewritten as:

$$H(C|W) = -\sum_{i=1}^{K} p(i|w) \log(p(i|w)) \qquad (5)$$

Where W is a random variable whose possibles values are strings of the Σ alphabet. We can calculate $H(C|W)$ for all individuals based on their genomes. This entropy may be expressed as the sum of entropies for each cluster as follows:

$$H(C|W)=\sum_{i=1}^{K} H(i|W) \qquad (6)$$

Where $H(i|W)$ is the entropy of cluster *i.* The idea is to minimize the entropy for each cluster. However, this fact involves a multi-objective optimization problem because minimizing the entropy of a cluster affects the entropy of any other. To resolve this problem we apply Pareto's Efficiency [18]. Our objective function may be written as:

$$min\left[H(1|W),H(2|W)\ldots H(K|W)\right] \qquad (7)$$

So, the GA must find the individuals that minimize this function which is represented as a vector of K dimensions. In what follows this vector is called *Entropy Vector*. Each individual has a Entropy Vector whose values are given by its genome. In order to determine the individual with the best vector, we apply the principle of *Pareto Dominance* [18]. The Pareto Dominance says that a Y vector dominates to Y^* if $\forall y_i \in Y, y_i \leq y_i^*$ and $\exists y_p$ such that $y_p < y_p^*$. In the context of VGA, a solution vector X of an Individual will dominate other solution vectors. The number of vectors dominated by X are called the *dominance value.* Thus, individuals with higher dominance value will be the best. The result of the evolutionary process yields a *Pareto Front*[18]. The fitness function for *i-th* individual (I_i) may be written as:

$$f(I_i)=dom_i \qquad (8)$$

Where dom_i is the dominance value of the i^{th} individual. However this function does not always assure that an individual with maximal dominance value is the best solution to the clustering problem. We, therefore propose a quality measure.

Our quality measure is based on the concept of *Mutual Information (MI)* [21]. It is a symetric measure that quantifies the mutual dependence between two random variables or the information that these share. In the context of our problem, the MI between two cluster u and v is given by:

$$I(u,v)=\sum_{i=1}^{R}\sum_{j=1}^{S} p(w_i,w_j)\log\frac{p(w_i,w_j)}{p(w_i)p(w_i)} \qquad (9)$$

where R and S are $|u|$ and $|v|$ respectively and $p(w_i, w_j)$ is the probability that the words w_i and w_j are similar. This probability is given by:

$$p(w_i,w_j)=\frac{|w_i \cap w_j|}{length(w_i)} \qquad (10)$$

where the intesection between two word is given by their common symbols. Clearly, all words of D' have the same length.

If $u \neq v$ then the value of $I(u,v)$ will be called *Mutual Information Intercluster (MI_{Inter})* . Otherwise this value will be called *Mutual Information Intracluster (MI_{Intra}).* A lower value of MI_{Inter} and higher value of MI_{Intra} means better clusters. So, we propose a quality measure given by:

$$Q=\frac{\sum_{i=1}^{K} MI_{Intra}(i,i)}{\sum_{i,j\leq K, i\neq j} MI_{Inter}(i,j)} \qquad (11)$$

An individual with higher value of Q means a better solution. Therefore the fitness function of the i^{th} individual may be defined as:

$$f(I_i)=dom_i Q_i \qquad (12)$$

However, we observe that an individual with a “good” fitness value does not always represent a global optimum. Thus, we assume that each individual must be subject to the following constraint : *The probability for all partition (cluster) of D must be greater than zero. Mathematically* $p(i)>0$ $\forall$ *i=1,2,..K (see Equation 2 and Equation 3).*

This constraint ensures that the individuals consists of non-empty clusters whose entropy is minimal. Otherwise the solutions will be outside of the feasible region. To encourage reproduction of feasible individuals (which represents feasible solutions) in every generation of VGA, we appeal to an penalty method [14] whose goal is to punish unfeasible individuals.

Here the penalty for unfeasible individual I_i is given by:

$$P(I_i)=J-\sum_{i=1}^{s}\frac{J}{m} \qquad (13)$$

where J is a large constant [$O(10^9)$], m is the number of constraints and s is the number of these which have been satisfied.

4 Numerical Experiment

In what follows we briefly describe how the test data set was generated. Subsequently we show several parameters and features of the performed tests. Finally we show the results. We call our proposal has been called *Entropic Evolutionary Clustering* (EEC).

4.1 The data set

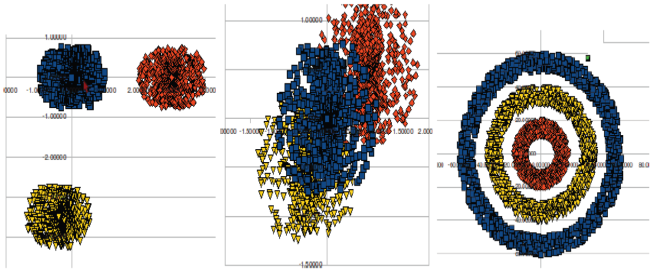
Three data sets are analyzed in this work. We shall call them "*A*", "*B*" and "*C*" respectively. Every set is composed of vectors (in a *3D* space) that belong to three different spheres which we call sphere 1, 2 and 3 respectively. There are 10,000 vectors in each one of the spheres. They were generated from.

$$x=x_0+r\sin\theta\cos\phi \qquad (14)$$

$$y=y_0+r\sin\theta\sin\phi \qquad (15)$$

$$z=z_0+r\cos\theta \qquad (16)$$

from uniformly distributed values for $r \in [0,1)$, $(0 \leq \phi \leq 2\pi$ and $0 \leq \theta \leq \pi)$. For set *A* the three centers of the spheres were chosen so that the spheres would not intersect. In set *B*, the chosen centers yield partially overlapping data. Finally, in set *C*, the spheres shared a common center. However, in the last set for sphere 1 $r \in [0,1)$; for sphere 2 $r \in [0, 0.666)$; for sphere 3 $r \in [0, 0.333)$. In this case, then, spheres 1, 2 and 3 share the same space where the density of 2 is larger than that of 1 and the density of 3 is larger than the other two. Our intent is to choose vectors in set A, B and C whose distribution is not uniform but Gaussian. To achieve this, we determined to divide the space of probabilities of a Gaussian curve in 20 equally spaced intervals. The area under the curve for a normal distribution with $\mu = 0$ and $\sigma = 1$ between -4 and +4 is very closely equal to one. Therefore, it is easy to see that 5%, of the observations will be between −4 and −1.654; 5%, will be between −1.654 and −1.280, etc. The required normal behavior may be approximated by selecting 50 of the uniformly distributed values from the interval $[-4,-1,654)$; another 50 from the interval $[-1.654,-1.280)$, etc. In all we will end up with 1000 vectors for every sphere. These vectors will now be very closely Gaussian. When data is normally distributed, a Bayesian classifier is optimal. The behavior of one such classifier will serve as a base point. To stress: when the distribution of the data set to classify is Gaussian, a Bayesian classifier yields the best theoretical results (by minimizing the probability of classification error independently of the degree of overlap between the distributions of the clusters) [7]. Hence, we resorted to Gaussian distributed data in order to establish a behavior relative to the best theoretical one when measuring the performance of non-traditional methods. Our claim is that, if the methods perform satisfactorily when faced with Gaussian data, they will also perform reasonably well when faced with other possible distributions. That is, we wish to show that the results obtained with non-traditional methods are close to those obtained with a Bayesian classifier for the same data set. This would mean that these results correspond to an efficient algorithm. The data sets are illustrated in Figure 6.



(a) Disjoint data set (b) Overlapped data set (c) Overlapped data set

Figure 6. Types of data set

5 Results

The values of the parameters of VGA are given in Table 4. These values were determined experimentally. As mentioned above we use the English Alphabet to transform the original data set. The VGA was run 20 times (with different seeds of the pseudo random number generator) per data set. The same data sets was tested with *K-Means* [22], *Kohonen Maps* [23] and *Fuzzy C-Means* [4]. Since it may be proven that a Bayesian Classifier is optimal when the data's pdf is Gaussian [7], we include a comparison with such a Bayesian Classifier. The results obtained with disjoint clusters are shown in Table 5. This allows us to see that the results of EEC are similar to those given by some alternative algorithms. The high effectiveness in all cases is due to the spatial distribution of data set. The results obtained with overlapping clusters are shown in Table 6 where we can see that the effectiveness decreases significantly in general.

Table 4: Parameters Test

Parameter Name	**Values**
N (Number of Individuals)	50
G (Generations)	4000
pm (Mutation probability)	0.00
pc (Crossover Probability) 0.99	0.99

However EEC showed better results than traditional methods and close results to Bayesian Classifier. The results obtained in the two last cases (overlapping and concentric clusters) are due to the fact that it is not possible to find a simple separable boundary. Therefore, the boundary decision is unclear and the vast majority of the clustering methods yield poor solutions. The closeness of the results obtained so far relative to a Bayesian Classifier, tells us that our approach is quite efficient. In future works we will report on experiments encompassing a wider range of data sets.

Table 5: Results obtained with disjoint clusters data set

Algorithm	**Average Effectiveness**
EEC	0.99
K-Means	0.98
Kohonen Maps	0.99
Fuzzy C-Means	0.98
Bayesian Classifier Effectiveness	0.99

Table 6: Results obtained with overlapping clusters data set

Algorithm	**Average Effectiveness**
EEC	0.87
K-Means	0.45
Kohonen Maps	0.72
Fuzzy C-Means	0.15
Bayesian Classifier Effectiveness	0.89

Table 7: Results obtained with concentric clusters data set

Algorithm	**Average Effectiveness**
EEC	0.71
K-Means	0.36
Kohonen Maps	0.38
Fuzzy C-Means	0.15
Bayesian Classifier Effectiveness	0.72

6 Conclusion

Following the minimum entropy principle we employed a genetic algorithm so that we were able to explore the solution space of the clustering problem. This approach resulted a better effectiveness with different data sets respect to *K-Means*, *Kohonen Maps* and *Fuzzy C-Means*. If we consider that Bayesian Classifier represents a theoretical limit then the most interesting result is the nearness of EEC respect this classifier. Our method promises to be a feasible alternative to find non-spherical clusters due to the results obtained with the concentric clusters of the data set C. However, we require testing several data sets that allow us to statistically ascertain that our method is good. We will report on these issues shortly. Additionally, data preprocessing proved to be a good alternative to reduce the computational complexity when the dimensionality of the data set is fairly high.

7 References

[1] Arndt, C., Information measures: information and its description in science and engineering, *Springer Verlag,* 2001

[2] Bäck, Th., Evolutionary Algorithms in Theory and Practice, *Oxford University Press*, 1996

[3] Cohen, J., Applied multiple regression/correlation analysis for the behavioral sciences, *Lawrence Erlbaum*, 2003

[4] Dunn, J. C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting CompactWell-Separated Clusters, *Journal of Cybernetics 3*, volume 3, 32–57, 1973

[5] Parzen E.: On the estimation of a probability density function and the mode. *Annals of Math. Stats.*, 33:1065-1076, 1962.

[6] Forrest, and Mitchell, What Makes a Problem Hard for a Genetic Algorithm? Some Anomalous Results and Their Explanation, MACHLEARN: *Machine Learning 13*, volume 13, 1993

[7] Haykin, Simon, *Neural networks: A comprehensive foundation*, MacMillan,1994

[8] Halkidi, Maria, Batistakis, Yannis, and Vazirgiannis, Michalis, On Clustering Validation Techniques, J. Intell. *Inf. Syst. 17(2-3)*, volume 17, 107–145, 2001

[9] Jenssen, R., Hild, KE, Erdogmus, D., Principe, J.C., and Eltoft, T., Clustering using Renyi's entropy, Neural Networks, 2003. *Proceedings of the International Joint Conference on*, volume 1, 523–528, 2003

[10] Joyce, James, *Bayes Theorem, The Stanford Encyclopedia of Philosophy*, Fall 2008 edition, Eds: Zalta, Edward N., 2008

[11] Kovacs, Ferenc, and Ivancsy, Renata, *A novel cluster validity index: variance of the nearest neighbor distance*, *WSEAS Transactions on Computers*, volume 3, 477-483, March 2006

[12] Kuri-Morales, Angel, A Methodology for the Statistical Characterization of Genetic Algorithms, MICAI, volume 2313, *Springer*, 79–88, Eds: Coello, Carlos A. Coello, de Albornoz, Alvaro, Sucar, Luis Enrique, and Battistutti, Osvaldo Cairó, 2002

[13] Kuri-Morales, Angel, and Aldana-Bobadilla, Edwin, Finding Irregularly Shaped Clusters Based on Entropy, ICDM, volume 6171, *Springer*, 57–70, Eds: Perner, Petra, 2010

[14] Kuri-Morales, Angel and Gutiérrez-García, Jesús, Penalty Function Methods for Constrained Optimization with Genetic Algorithms: A Statistical Analysis, *MICAI, volume 2313, Springer*, 108–117, Eds: Coello, Carlos A. Coello, de Albornoz, Alvaro, Sucar, Luis Enrique, and Battistutti, Osvaldo Cairó, 2002

[15] Lee, Y., and Choi, S., Minimum entropy, k-means, spectral clustering, Neural Networks, 2004. *Proceedings IEEE International Joint Conference on*, volume 1, 2005.

[16] Li, H., Zhang, K., and Jiang, T., Minimum entropy clustering and applications to gene expression analysis, *IEEE Computer Society*, 2004.

[17] Pearson, K., Principal components analysis, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6(2)*, volume 6, 559, 1901

[18] Podinovskii, VV, and Nogin, VD, Pareto-Optimal Solutions of Multicriteria Problems, *Nauka*, Moscow, 1982

[19] Rudolph, G., Convergence Analysis of Canonical Genetic Algorithms, *IEEE Transactions on Neural Networks 5(1)*, volume 5, 96–101, January 1994

[20] Shannon, C. E., and Weaver, W., The Mathematical Theory of Communication, *Scientific American*, July 1949

[21] Vinh, N.X., Epps, J., and Bailey, J., Information theoretic measures for clusterings comparison: is a correction for chance necessary?, *Proceedings of the 26th Annual International Conference on Machine Learning*, 1073–1080, 2009

[22] McQueen, J. B., Some Methods of Classification and Analysis of Multivariate Observations, *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297, Eds: Cam, L. M. Le, and Neyman, J., 1967

[23] Kohonen Teuvo, Self-organizing maps, *Springer-Verlag,* New York, Inc., 199