Isolation of a Sequence Homolog to More Axillary Branches MAX2 Gene in *Hibiscus rosa-sinensis* and its Use as Genetic Marker

[1]CNR-IBBA Institute of Agricultural Biology and Biotechnology, Milano, Italy

[2]Consiglio per la Ricerca e la Sperimentazione in Agricoltura - Centro di Ricerca per le Colture Industriali (CRA-CIN), Bologna, Italy

[3]Consiglio per la Ricerca e la Sperimentazione in Agricoltura - Unità di Ricerca per la Floricoltura e le Specie Ornamentali (CRA-FSO), Corso Inglesi, Sanremo (Imperia), Italy

[4]University of Modena and Reggio Emilia, Department of Agricultural and Food Sciences, Reggio Emilia, Italy

Abstract

Lateral branching plays an important role in the elaboration of adult plants architecture. Herein, we adopted a modified AFLP approach combined with a degenerate primer amplification to identify and isolate in the underinvestigated ornamental species H. rosa-sinensis an orthologous element of the MAX2 gene (More Axillary Branches), which acts downstream of the branching inhibition signaling pathway. A specific gene fragment was cloned and sequenced from nineteen H. rosa-sinensis cultivars and twelve Hibiscus botanical species and different significant nucleotide polymorphisms among genotypes that were observed. The comparative analysis revealed a high conservation of DNA sequences among cultivars and wild species sexually compatible with H. rosa-sinensis. The deduced amino acid sequences of the Hibiscus isolated fragments reveal four characteristic repeat regions showing high identity with other F-box/Leucine Reach Repeat MAX2 homologous sequences. The cloned fragment is a likely candidate gene to be validated for association with phenotype to release a gene-derived "perfect marker" for the compact habit trait.

Keywords:

Hibiscus - AFLP approach - ornamentals - shoot branching.

Introduction

Hibiscus rosa-sinensis is one of the most widely planted ornamental shrubs cultivated throughout the tropics and sub-tropics. Numerous varieties and hybrids are particularly appreciated in garden and landscape for their vigorous growth habit, but they are mainly employed as a pot plant when growth retardants for keeping a reduced plant size are applied. A compact basal branching growth habit is generally preferred.

Lateral branching structures exist in many forms throughout higher plants and even if plant architectures are influenced by environmental factors, their species-specific characteristics indicate the presence of widely conserved genetic regulatory mechanisms (McSteen and Leyser 2005, Johnson et al 2006).

However, the involvement of a novel hormone-signaling pathway in the regulation of bud growth has been inferred by genetic analyses of mutants that have enhanced shoot branching phenotype in Arabidopsis, pea, petunia and rice (Booker et al 2005, Morris et al 2001, Snowden and Napoli 2003, Arite et al 2009). In *Arabidopsis thaliana*, a suite of mutants with More Axillary Branches encoded by In different plant species, pre-existing axillary meristems may either lie dormant for long periods or they may develop into branches instantaneously. This bud growth can be activated by intrinsic factors, and hormones play a crucial role in shoot branching control (Leyser 2009). It has been known for many years that auxin synthesize in the apex inhibits axillary meristems outgrowth, whereas cytokinin promote it efficiently by regulating the shoot branching phenomena (Liang et al 2010).

the MAX1, 2, 3, and 4 genes has been analyzed by a combination of grafting and molecular techniques (Bennet et al 2006). The recessive mutations (*max*) cause premature and enhanced outgrowth of lateral shoots in combination with modest pleiotropic effects. These studies suggested

that MAX1, 3 and 4 are involved in the synthesis of a mobile signal, whereas the MAX2 gene product mediates perception and response to the signal (Leyser 2009). Particularly, MAX2 has been shown to encode a nuclear localized F-box leucine-rich repeat (LRR) protein within the SCF (Skp1-Cullin-F box) complex that catalyzes the ubiquitination of proteins, and thus target them for proteasomal degradation (Xu et al 2009). In the case of MAX2, one or more proteins that activate bud growth are in the wild-type targeted for destruction by the MAX2 F-box LRR product (Stirnberg et al 2007). Presumably, these proteins, which would be stabilized in the absence of MAX2 activity, would in some way promote branching.

More recently, Wang et al (2013) demonstrated in Arabidopsis that the strigolactone hormone inhibits auxin transport, suggesting a complex interaction between these two hormones and the MAX2 F-box binding site in the protein degradation system.

Given the metabolic complexity of plants, there are probably more, perhaps many more, small molecules with signaling function. However, the discovery of regulatory mechanisms promoting the axillary branch proliferation could provide an environmentally independent, rapid and helpful tool for preliminary screening of genotypes characterized by a compact basal branching growth habit, suitable for pot plant cultivation.

With the aim to identify a gene-derived 'perfect' molecular marker associated to the compact plant architecture, we isolated conserved sequences for the MAX2 gene in the underinvestigated ornamental species *Hibiscus rosa-sinensis.* The knowledge gained through the previous AFLP characterization of a collection of *H. rosa-sinensis* cultivars (Braglia et al 2010) allowed us to develop a new strategy. Starting from plant MAX2 gene sequences, we followed a combined approach of degenerate primer PCR together with AFLP technique

According to the full-length cDNA sequence of Arabidopsis MAX2 gene (NM_129823), pea RAMOSUS4 gene (DQ403159), rice LRR-repeat MAX2 homolog (*Oryza sativa Japonica* group) (Q5VMP0) and poplar F-box family protein mRNA sequence (*Populus trichocarpa*) (XM_002320376) showing strong homology at the amino acid level, a consensus sequence by multiple sequence alignment was generated. A set of six degenerate primers were then designed and tested on three different *Hibiscus rosa-sinensis* genomic DNAs. PCR reactions were performed in a 50 µl containing 1X PCR Buffer (HotStartTaq®Plus Buffer Qiagen, Germany), 0.2 mM each dNTP, 2 mM $MgCl_2$, 150 ng of DNA template, 1.6 µM primer and

2.5U Taq DNA Polymerase (HotStartTaq®Plus Qiagen, Germany). The PCR conditions were 2 min at 94° C, 5 cycles 30 s at 94° C, 45s at 48° C, 2 min at 72° C, followed by 35 additional cycles 30 s at 94° C, 45 s at 58° C, 90 s at 72°. The reactions were held at 4°C after a final extension at 72°C for 10 min.

CCYTGRAAGTGNCCNAGCTT-3') yielded the expected size fragment. This PCR product was subsequently cloned (TA Cloning® kit, Invitrogen) and sequenced, then analyzed using bioinformatic tools at the websites
http://www.ebi.ac.uk/Tools/ and http://www.ncbi.nlm.nih.gov/.

Hibiscus specific primers (notable as Hsp_, Table 1) were designed on the isolated fragment sequence in forward and reverse. An AFLP (Amplified Fragment Length Polymorphism)-based approach was adopted to extend the *Hibiscus* DNA segment outside the boundary known sequence. Restricted/ligated fragments (EcoRI/MseI), hereafter R/L were generated according to the AFLP protocol reported by Vos et al (1995) from 300 ng of genomic DNA. The obtained R/L products were used to test different Hsp_ primers in combination with AFLP primers (Table 1). These latter, named Ead_pr and Mad_pr, had the 5'-region complementary to the adapter and the restriction site sequence without selective nucleotides at the 3'-end.

PCR reactions were performed in a 25 µl

containing 1X PCR Buffer (HotStartTaq®Plus Buffer Qiagen, Germany), 1 µl R/L, 0.2 mM each dNTP, 1.5 mM $MgCl_2$, 0.5 µM for the Hsp_ primer and 0.1 µM for the AFLP primer with 2.5U of Taq DNA Polymerase (HotStartTaq®Plus Qiagen, Germany). The following PCR conditions were used: 5 min at 95 °C, 13 cycles of 30 s at 90°C, 30 s at 67°C, 60 s at 72°C with a decrease of 0.7°C of the annealing temperature carried out in each cycle followed by 27 additional cycles of 30 s at 90° C, 30 s at 56° C, 60 s at 72°. The reactions were held at 4°C after a final extension at 72°C for 10 min. The obtained PCR product was subsequently cloned (TA Cloning® kit, Invitrogen) and sequenced. One hundred additional base pairs were achieved from the AFLP-based approach and two new Hsp_ reverse primers were synthesized (Fig. 1, Table 1).

Fig. 1: *Hibiscus* MAX2 Gene Fragment Schematic Representation. MseI Restriction Site is Reported at the 5' Region. Black Arrows Indicate the Name and Position of Hsp_ Specific Primers.

Table 1: *Hibiscus* Specific Primers and AFLP Core Primers

Name	Primer type	Primer sequence (5'→3')
Hsp_Up1	Forward specific primer	AACCACCGCCGCTTGTCCTAAC
Hsp_Up2	Forward specific primer	AGACGAGACCTTGTTGGCAGTGG
Hsp_Dw1	Reverse specific primer	AGCTTCAGCACTCTCAAATCCTT
Hsp_Dw2	Reverse specific primer	CAACGGCATCCTCGGAAGTAAAC
Hsp_Dw05	Reverse specific primer	GTAATAGACAGCTCTCGCAG
Ead_pr	EcoRI primer	GACTGCGTACCAATTC
Mad_pr	MseI primer	GATGAGTCCTGAGTAA

The selected primer pair Hsp_Up1\Hsp_Dw1 was tested on all genomic samples using the same PCR conditions reported above. Amplified fragments were sequenced to assess nucleotide polymorphisms.

The amino acid sequences were deduced and the sequence comparison was conducted through database search using UniProt (Universal Protein Research http://www.uniprot.org)

Results

The degenerate primer amplification allowed the identification of a DNA fragment (~350 bp) showing 65% similarity to the Arabidopsis MAX2 gene and the deduced amino acid sequence revealed a high degree of homology with those of other F-box subunit proteins from various biological sources: *Populus trichocarpa* (64%), *Arabidopsis thaliana*

(64%), *Pisum sativum* (61%) and *Oryza sativa* (32%), most of them employed in the degenerate primer design.

Concerning to the AFLP approach, all tested specific primers in combination with the primer Ead_pr did not produce any amplification products (data not shown). Whereas, the primer combination Mad_pr/Hsp_Up1 allowed to extend the *Hibiscus* MAX2-like sequence downstream the 3' boundary known sequence of one hundred additional nucleotide base pairs. Unfortunately, the presence of a *Mse*I restriction site at the 5' region of the isolated fragment (named HibMAX2-like) did not permit to extend the sequence upstream (Fig.1).

The HibMAX2-like sequences could be amplified in all cultivars of *Hibiscus rosa-sinensis* and *Hibiscus* species except for *H. cannabinus* (kenaf). Control reactions on genomic DNA samples of Arabidopsis and pea did not produce any amplicons (Data not shown). The *Hibiscus* MAX2-like nucleotide sequences were submitted to NCBI database with the Accession Numbers JF813799-JF8137824. A neighbour-joining tree was then constructed (Fig. 2) for all *Hibiscus* isolated nucleotide sequences. In

this tree, a main cluster (A, bootstrap value 98%) could be recognized, grouping all *Hibiscus rosa-sinensis* cultivars. Furthermore, some botanical species such as *H. arnottianus, H. boryanus, H. denisonii, H. genevii, H. kokio*, *H. schizopetalus* and *H. storckii* were spread throughout the main cluster. Conversely, *H. calyphyllus, H. moscheutos, H. panduriformis, H. syriacus and H. tiliaceus,* were grouped in separate clusters.

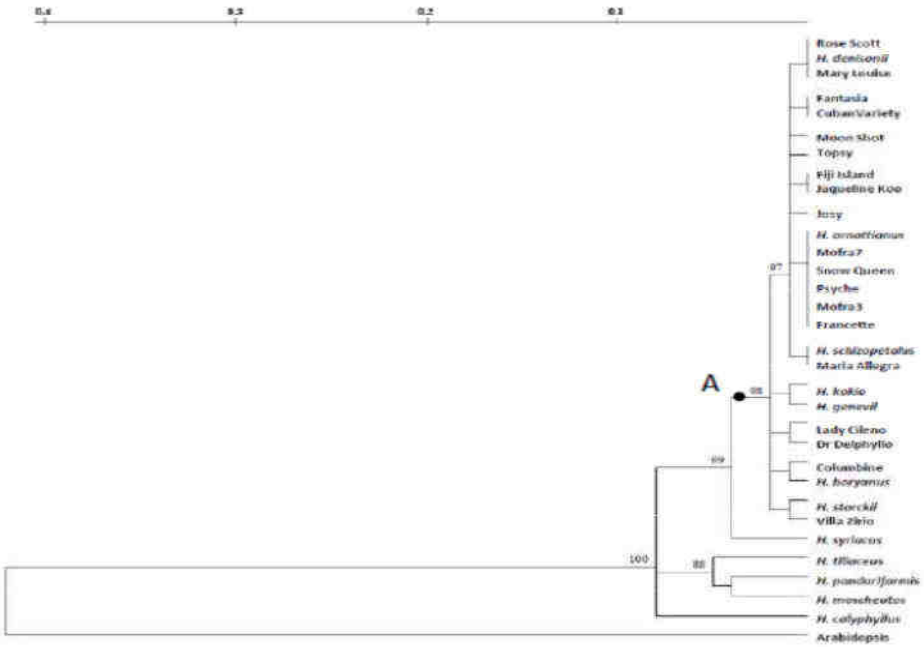
Fig. 2: Neighbor-joining Tree Built from HibMAX2 Nucleotide Sequences of 19 *Hibiscus rosa-sinensis* Varieties and 12 *Hibiscus* Botanical Species. The Arabidopsis MAX2 Sequence was Reported as an Outgroup. Numbers on Nodes Indicate the Bootstrap Values after 1000 Replicates.

Although the deduced HibMAX2-like amino acid sequences lack the conserved N-terminal and C-terminal domains, the comparison of these sequences revealed the presence of the characteristic repeat regions in all samples analyzed (black

boxes in Fig. 3). Indeed, *in silico* analysis showed the presence of four LRRs of the motif LxxLxL, with L (leucine), sometimes replaced by other aliphatic residues: valine, isoleucine and phenylalanine.

Material and Methods

Genomic DNA was isolated from nineteen *Hibiscus rosa-sinensis* cultivars with different and contrasting plant architectures and thirteen *Hibiscus* botanical species *(H. arnottianus* G., *H*. *boryanus H*. and A,*. H. calyphyllus* Cav., *H. cannabinus* L., *H. denisonii* B., *H. genevii* B., *H. kokio* H., *H. moscheutos* L., *H. panduriformis* B., *H. schizopetalus* H., *H. storckii* S., *H. syriacus* L., *H. tiliaceus* L.), selected from materials collected at the CRA-FSO in Sanremo (Italy). DNA from one hundred milligrams of fresh leaves was extracted using the DNeasy Plant Mini Kit (Qiagen, Germany) following the modified protocol reported for *Hibiscus* spp. by Braglia et al (2010).



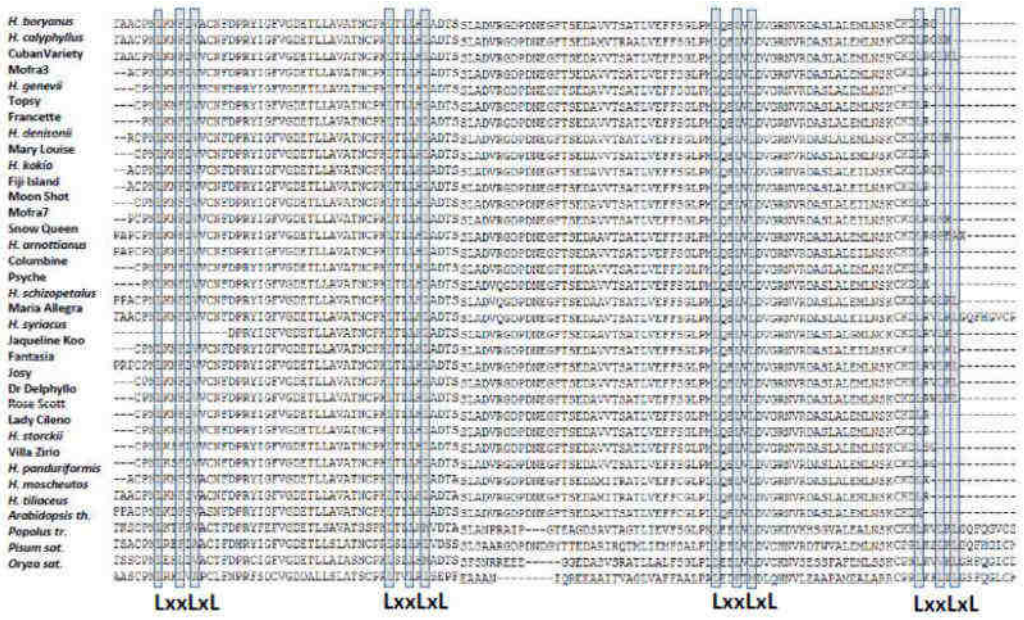


Fig. 3: Comparison of the Putative Amino Acid Sequences of HibMAX2 Obtained from Cultivars and Botanical Species. The Homologous Sequences of Arabidopsis, Pea, Rice and Poplar are Compared with HibMAX2. The Imperfect LRR Repeats are Shaded.

Discussion

The combined approach proposed here has proved its usefulness for allowing the rapid cloning and characterization of specific conserved genes, as well for providing plant genomic fingerprinting information. An orthologous element for the Arabidopsis More Axillary Branching gene was successfully cloned and sequenced, for the first time, from several *H. rosa-sinensis* cultivars and from *Hibiscus* wild species. This general approach is particularly useful when dealing with plant species for which no or poor information is available at the genomic level. Minor changes could be made to the suggested protocol, *i.e.* the choice of the restriction enzyme could likely increase the efficiency of the method.

The high conservation sequence degree observed among commercial varieties and the *Hibiscus* species sexually compatible with *H. rosa-sinensis* are in agreement with the cytogenetic evidence produced by Singh and Khoshoo (1989), which showed that these inter-fertile species have contributed to the extensive genetic
variability of *H. rosa-sinensis*. The revealed HibMAX2-like sequence analysis is consistent with secondary ranks of taxa (Sections) proposed by Pfeil and Crisp (2005) through chloroplast DNA analysis. In fact, the highest similarity values for the target sequence were achieved among the analyzed *H. rosa-sinensis* cultivars and the sexually compatible species, all belonging to the Lilibiscus Section, while the lower values were observed for species of different taxonomic Sections such as *H. syriacus, H. panduriformis, H. moscheutos, H. tiliaceus* and *H. calyphyllus*. Concerning the *H. cannabinus,* it could likely either do not possess the HibMAX2 gene, or possess a highly differentiated element, therefore not amplified. Kenaf is one of the fast-growing plants classified in the Furcaria Section of *Hibiscus*; it has both annual and biennial genotypes, often not branched. This differentiates from the other examined species (shrubs or small trees)

characterized by perennial life cycle, with complex vegetative morphologies (Craven et al 2003). Moreover, a previous study (Braglia et al 2010) had revealed the lowest genetic similarity value between *H.*

rosa- sinensis cultivars and kenaf defining this latter as the most distantly related species within the *Hibiscus* genus.

The occurrence of LRRs in the *Hibiscus* isolated fragment assigns this sequence among the F-box genes, one of the largest multigene superfamilies involved in shoot lateral branching growth. Members of this protein family function as subunits of the

multiprotein Skp-Cullin-Fbox for polyubiquitination and degradation by the 26S proteasome (Xu et al 2009). In particular, the F-box LRR proteins confer substrate specificity to the SCF complex via their two distinct functional domains: the first domain (F-box) binds to another subunit of the SCF complex, the second domain (LRR repeats) interacts with specific proteins to be polyubitquinated (Stirnberg et al 2007).

Although the present results are the first step in the isolation of the whole *Hibiscus* specific element for MAX2 gene, the cloned fragment can be already investigated for association to the branched trait, to evaluate its utility in marker-assisted breeding schemes. Further studies including the isolation of a cDNA fragment (working back to the full length through RACE-PCR technique), the mRNA expression analysis and the functional variant identification are in turn necessary to better characterize the HibMAX2-like sequence, as well as to clarify its involvement in the axillary branch proliferation mechanisms.

Acknowledgements

We wish to thank Mr. Cesare Bianchini and Dr. Marco Ballardini for their support in managing the germplasm collection. Research founded by the Italian Ministry of Agriculture in the framework of the project “Risorse tecniche e genetiche per il florovivaismo (FLORIS)”.

References

1. Arite, T., Umehara, M., Ishikawa, S., Hanada, A., Maekawa, M., Yamaguchi, S. & Kyozuka, J. (2009). "D14, a Strigolactone-

2. Bennett, T., Sieberer, T., Willett, B., Booker, J., Luschnig, C. & Leyser, O. (2006). "The Arabidopsis MAX Pathway Controls Shoot Branching by Regulating Auxin Transport," *Current Biology*, 16 553–563.

3.Booker, J., Sieberer, T., Wright, W., Williamson, L., Willett, B., Stirnberg, P., Turnbull, C., Srinivasan, M., Goddard, P. & Leyser, O. (2005). "MAX1 Encodes a Cytochrome P450 Family Member that Acts Downstream of MAX3/4 to Produce a
Carotenoid-Derived Branch-Inhibiting Hormone," *Developmental Cell*, 8 443–449.

4.Braglia, L., Bruna, S., Lanteri, S., Mercuri, A. & Portis, E. (2010) "An AFLP-based assessment of the genetic Diversity within Hibiscus rosa-sinensis and its Place within the Hibiscus Genus Complex," *Scientia Horticulturae*, 123 (3) 372-378.

5.Craven, L. A., Wilson, F. D. & Fryxell, P. A. (2003). "A Taxonomic Review of Hibiscus Sect. Furcaria (Malvaceae) in Western Australia and the Northern Territory," *Australian Systematic Botany*, 16 (2) 185-218.

6.Hillis, D. M. & Bull, J. J. (1993). "An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis," *Systematic Biology*, 42 182-192.

7.Johnson, X., Brcich, T., Dun, E. A., Goussot, M., Haurogné, K., Beveridge, C. A. & Rameau, C. (2006). "Branching Genes are Conserved Across Species. Genes Controlling a Novel Signal in Pea are Co Regulated by Other Long-Distance Signals," *Plant Physiology*, 142 1014–1026.

8.Leyser, O. (2009). "The Control of ShootBranching: An Exampleof Plant Information Processing," Plant, Cell &*Environment* 32, 694–703.

9.Liang, J., Zhao, L., Challis, R. & Leyser, O.(2010). "Strigolactone Regulation of Shoot

International Journal of Artificial Intelligence and Soft Computing (IJAISC) Vol. 1, No. 1

Multiobjective Partitional Clustering for Fuzzy and Mixed data through Hill Climbing

Pablo Barbaro Martinez Pedroso,Hotel Melia Cayo Coco,Ciego de Ávila, Cuba,+53 58053324

DarianHoracio Grass Boada,Centro de Aplicaciones de Tecnologías Avanzadas (CENATAV),La Habana, Cuba,+53 54500567

Amanda Robinson,Provalis Research,Montreal Canada.

A*BSTRACT*

In this paper we have designed and implemented multiplepossible stochastic hill climbing alternatives, applied to mixed (continuous and categorical) data in a fuzzy context, as proposed solutions to a multiobjective partitional clustering problem. To validate the efficacy of this approach we selected the external validity indexes Adjusted Rand Index (ARI) and Minkowski Score (MS). An approach capable of performing multiobjective partitional clustering with mixed data, which also provides solutions modeled with fuzzy logic, allowing for a better description of the distribution of objects among the clusters, was obtained as a result of the research.

CCS Concepts

• Theory of computation→ Unsupervised learning and clustering • Applied computing→ Multi - criterion optimization and decision-making• Computing methodologies→ Cluster analysis.

Keywords

Partitional clustering; multiobjective hill climbing ;fuzzy domain; mixed data.

1. INTRODUCTION

The large volume of information stored in enterprises, entities, institutions, etc. surpasses human capability of analyzing data analysis and comprehension. Knowledge discovery from databases process can be applied in order to extract unknown and interesting trends. Partitional clustering is a relevant unsupervised task of this process, which is defined as follows

Giving a set of objects $X = \{x_1, x_2, \dots, x_n\}$ where $x_j = (x_1, x_2, \dots, x_d) \in R^d$ and x_{ji} is a feature of the object. X is divided into k partitions (clusters, groups) $C = (C_1, C_2, \dots, C_k),\ (k \leq n)$ where:

$$C_i \neq \phi, i = 1, \dots, k \quad (1)$$

$$U_{i=1}^{k} C_i = X \quad (2)$$

$$C_i \cap C_j = \phi \quad i, j = 1, \dots k, i \neq j \quad (3)$$

Equation (3) defines crisp partitional clustering. However, there exist domains where frontiers of groups are not very clear. Modeling the problem as fuzzy partitional clustering allows more accuracy in respect to memberships of objects among the groups, in contrast to crisp partitional clustering that considers complete belongingness of objects to clusters. This information plays an important role for the decision maker. In order to model fuzzy partitional clustering, Equation (3) is substituted for a data structure known as a membership matrix, defined in (Xu 2009) as:

$U = \left[u_{ji}\right]_{N \times c}$ where $u_{ji} \in [0,1]$ is membership coefficient of j-thgroup. Satisfying the following restrictions:

$$\sum_{i=1}^{c} u_{ji} = 1, \forall j \quad (4)$$

$$0 < \sum_{j=1}^{N} u_{ji} < N, \forall i \quad (5)$$

Where c is the number of clusters and N the total amount of objects. Equation (4) is the membership distribution of each object among the clusters whereas Equation (5) prevents obtaining empty groups.

As (Xu 2009) outlines, optimal partitioning cannot be obtained due to the extreme computational cost. Thus heuristics are needed, although optimal solutions cannot be provided, at least near-optimal solutions can be found.

A well known technique is the k-means algorithm, the procedure is as follows. First k objects are randomly selected as centers of clusters. All other objects are grouped to the nearest center, based on distance metric (Euclidian distance). Then centers of clusters are updated through Equation (7). This procedure iterates

until no new centersare computed or an iteration limit is reached. The final centers obtained represent and characterize the clusters.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} (|p - z_i|)^2 \qquad (6)$$

$$m_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j \qquad (7)$$

The K-means algorithm is only suitable for a numeric domain since Euclidian distance is purely numeric. Nevertheless, many real life data sets are categorical in nature as is pointed in (Anirban Mukhopadhyay 2007), and a variation of k-means is needed.

In this variation dissimilarity between objects is measured by Equation (8) extracted from (Huang 1998):

$$d(X, Y) = \sum_{j=1}^{r} \delta\left(x_j, y_j\right) \qquad (8)$$

where:

$$\delta(x_j, y_j) = \begin{cases} 0 \ (x_j = y_j) \\ 1 \ (x_j \neq y_j) \end{cases} \qquad (9)$$

Centers are constructed with the mode of each feature of cluster objects, known as k-modes.

However, most of the real problemed data sets are mixed in nature (numeric and categorical features). In such domains none of the previous alternatives could by applied in their original design. To overcome this limitation (Huang 1998) proposes an integration of both techniques in a procedure called k-prototypes. In this procedure distance function Equation (6) and dissimilarity function Equation (8) are used to compare numerical and categorical features respectively. Thus the difference between two objects X and Y with mixed features denoted as vector of attributes as $A_1^r, A_2^r, \dots, A_p^r, A_{p+1}^r, \dots, A_m^c$ is calculated as follows:

$$d^2(X, Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{p+1}^{m} \delta(x_j, y_j) \qquad (10)$$

Where γ is used to avoid favoritism of any features types. Computing representative object Equation (7) is used for numerical data and mode for categorical data.

Representative objects of clusters can be observed being constructed. In (Kamber 2006) authors state that k-means and variations as previously presented, are sensitive to outliers, i.e. extremely distant data from a cluster can degrade substantially the solution. An alternative is selecting an existing object as a representative and all other objects are grouped to the most similar, computed with Equation (11).

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j| \qquad (11)$$

Where E is total sum of error, p an object of cluster C_j and o_j its correspondent representative object. This strategy is known as k-medoids, a medoid being the most centrally located object in a cluster, i.e. representative object.

Methods based in k-medoids are not restricted to specific data type, thus distant metric Equation (6) or dissimilarity measure Equation (8) can be adapted without limitation to k-medoid procedure. Nevertheless Equation (10) is selected in order to cover a mixed data domain. So far a mixed data domain has been tackled however most of the real data set does not have clear enough frontiers between clusters. In order to overcome this limitation, a variation of the previously mentioned technique is adopted, known as fuzzy k-medoids, it partitions the entire data set into k clusters considering that each object belongs to all clusters with a degree of belongingness or membership, defined in (Mukhopadhyay 2013)as follows.

$$J_m(U, Z: X) = \sum_{j=1}^{n} \sum_{i=1}^{k} u_{ji}^{m} * d^2(x_j, Z_i) \qquad (12)$$

Where $U = [u_{ji}]$ represents the matrix of fuzzy partition, u_{ji} membership degree of object j to cluster i and $Z = \{z_1, \dots, z_k\}$ vector of medoids.

$$u_{ji} = \frac{1}{\sum_{p=1}^{k} \left(\frac{d(Z_i, m_j)}{d(Z_p, m_j)} \right)} \qquad (13)$$

The method starts with k randomly selected medoids. In each iteration, after membership matrix is calculated with Equation (13), it is used to re-computmedoids with Equation (14). Medoid$Z_i$of i-th cluster satisfies $z_i = x_p$ such as:

$$p = \boldsymbol{argmin}_{1 \le j \le n} \sum_{k=1}^{n} {u_{ik}}^{m} * d(x_j, x_k) \qquad (14)$$

The techniques presented so far optimize only one criterion (compactness) over the entire data set. For this type of distribution optimizing compactness can yield good solutions. However, since clustering is an unsupervised technique there is no previous knowledge about distribution of the objects. Moreover, one criterion alone cannot uncover groups of distinct types, therefore, and as (Hruschka 2009)suggests quality of clusters should be measured by multiple criteria instead of a single criterion.

Optimizing more than one criterion has been proposed in two main approaches ensemble and multiobjective (Hruschka 2009).(Handl J. 2007) and (Hruschka 2009) outline of ensemble which tends to be more robust and provides better solutions than single objective optimization, they posit that it does not exploits entirely the potential of using various criteria. Since ensemble is restricted to integrating solutions provided by multiples single objective optimization techniques it does not exploit solutions that are simultaneously optimized. On the other hand such solutions are explored by the multiobjective approach. The multiobjective approach introduced in (Handl J. 2004a), optimizes simultaneously various objectives, conflictive between them, thus optimizing one, degrades other. In such an approach, many objective functions are considered as the problem, and every one with the same level of priority.
The formalization of the Multiobjective optimization problem is extracted from (Mukhopadhyay 2007a):
Find the vector $\bar{x}^* = [x_1^*, x_2^*, \dots, x_n^*]^T$ of decision variables that will satisfy the m inequality constraints

$$g_i(\bar{x}) \geq 0, \qquad i = 1,2, \dots, m \qquad (15)$$

the$p$ equality constraints

$$h_i(\bar{x}) = 0, \qquad i = 1,2, \dots, p \qquad (16)$$

and optimizes the vector function

$$\bar{f}(\bar{x}) = \left[\bar{f}_1(\bar{x}), \bar{f}_2(\bar{x}), \dots, \bar{f}_k(\bar{x})\right]^T \qquad (17)$$

To clarify when a solution is considered optimal principles of Pareto are applied in this research. Related concepts can be found in (Coello Coello 2007) and are defined as follows.

A solution $x \in \Omega$, is said to be optimal of Pareto respecting to Ω if and only if there is no $x' \in \Omega$ for which $v = F(x') = \left(F_1(x'), \dots, F_r(x')\right)$ dominates $u = F(x) = (F_1(x), \dots, F_r(x))$. A vector $u = (u_1, \dots, u_k)$ dominates another vector $v = (v_1, \dots, v_k)$ denoted by $(u \preccurlyeq v)$ if and only if u is partially less than v, this is, $\forall i \in \{1, \dots, r\}, u_i \leq v_i \wedge \exists i\{1, \dots r\}$, such as $u_i < v_i$. Applying principles of Pareto to MOO rather than one, a set of solutions is obtained, known as the Pareto optimal set, which is in fact the aim of the process. For a given MOO problem, $F(x)$, Pareto optimal setP^*, is defined as:

$$P^* := \{x \in \Omega \mid \nexists\, x' \in \Omega\, F(x') \preccurlyeq F(x)\} \qquad (18)$$

The objective of this paper is to develop a multiobjective optimization procedure for partitional clustering capable of covering mixed and fuzzy data.

1.1 Related Work

Partitional clustering is an NP-hard problem (Dutta 2012b). In absence of the exact solution, metaheuristics provide near optimal solutions in a reasonable response time.

Much effort has been exertedwith the perspective that evolutionary algorithms (EAs) provide good solutions to multiobjective partitional clustering. Approaches to multiobjective partitional clustering based on evolutionary algorithms can be found in (Mukhopadhyay 2007a), (Bandyopadhyay 2007),

(Mukhopadhyay 2010), (Handl J. 2004a), (Handl J. 2005), (Handl J. 2007), (Dutta 2012a), (Dutta 2012b), (Dutta 2012c), (Dutta 2012d), (Dutta 2013) for crisp partitions. Whereas for fuzzy partitions, it can be found in (Mukhopadhyay 2007b), (Mukhopadhyay 2009), (Saha 2011), (Mukhopadhyay 2013), (Saha 2013a).

On the other hand, local search metaheuristics, have not been sufficiently exploited for multiobjective partitional clustering problems, as was pointed out in (Bandyopadhyay 2008), because of its nature of searching in the neighborhood of a single solution in each iteration. However, different effective methods have been developed, such as (Smith and Misra 2005), (Bandyopadhyay 2008) and (Saha 2009), (Saha 2013b) based in simulated annealing (SA) whereas tabu search (TS) was used in (Beausoleil 2007) and as a hybrid component in (Caballero 2008) and (Caballero 2009).

In the surveyed literature neither random search (RS) nor hill climbing (HC) had been used as searching strategy to tackle the multiobjective partitional clustering problem. Nevertheless previous works had implemented local search for multiobjective optimization. In (Infante Abreu 2014)tabu search, simulated annealing and hill climbing were used. Experimental results demonstrated a good and stable performance of hill climbing over simulated annealing and tabu search.

In the research presented by (Díaz Pando and Rosete Suárez 2013)the multiobjective optimization problem was tackled with various methods (hill climbing, random search, tabu search, simulated annealing and genetic algorithm). As a result they provide evidence of perceptible superiority of hill climbing over the rest of techniques used, with respect to average of convergence. Finally (Díaz 2001) carried out multiples experiments to compare hill climbing, restart hill climbing and genetic algorithm in a multiobjective optimization problem. Based on the data set used, such research concludes hill climbing is as good or better than genetic algorithm and restart hill climbing overcomes genetic algorithm performance.

In the light of this, and based on No Free Launch theorem, hill climbing is selected as searching strategy for the multiobjective partitional clustering problem.

2. MULTIOBJECTIVE PARTITIONAL CLUSTERING – HILL CLIMBING

2.1 Fitness computation

In (Handl J. 2007) objective functions are classified (depending on what type of distribution present clusters identify) into three categories: compactness, connectedness and spatial separation. Compactness tries to find clusters where objects are very similar to centers, whereas connectedness looks for convex structures and spatial separation delimit as much as possible frontiers between clusters.

The literature surveyed presents that in crispclustering case, compactness and connectedness are more often the selected criteria(Handl J. 2004b), (Handl J. 2004a), (Handl J. 2005), (Handl J. 2007), (Matake 2007),

(Chun-Wei 2012) and (Saha 2013b) compared with compactness and spatial separation (Dutta 2012a), (Dutta 2012b), (Dutta 2012c) and (Mukhopadhyay 2007a).

All papers that deal with fuzzy data (Suresh 2009), (Mukhopadhyay 2010), (Di Nuovo 2007), (Saha 2011), (Mukhopadhyay 2009), (Mukhopadhyay 2007b), (Bandyopadhyay 2007), (Saha 2013a) and (Mukhopadhyay 2013),except for (Di Nuovo 2007), measure compactness and spatial separation simultaneously. As can be observed, compactness is always measured independently of crisp/fuzzy clustering and numerical/categorical data. This criterion describes in principle the basis of clustering, i.e. identify clusters where objects are very similar to its center. Another interesting aspect is that for crisp partitioning, compactness is simultaneously optimized with connectedness, in contrast with fuzzy partitioning where spatial separation is used.

In the light of this, fitness computation is oriented to measure compactness and spatial separation. Therefore we selected J_m Equation (12) and *Xie-Beni* Equation (20) exactly as were used in (Suresh 2009), (Bandyopadhyay 2007) and (Mukhopadhyay 2013).

Function *Xie-Beni* (XB) measures average between overall deviation σ, (which is in fact J_m) and minimal separation sep of the clusters (Beni 1991).

$$\sigma = \sum_{j=1}^{n}\sum_{i=1}^{k} u_{ij}^2 * \ d^2\left(\overrightarrow{x_j}, \overrightarrow{Z_\iota}\right) \qquad (18)$$

$$\boldsymbol{sep} = \ min_{i\neq j}\{d^2(Z_i, Z_j)\} \qquad (19)$$

$$X\mathrm{B} \ = \frac{\sigma}{n*sep} \qquad (20)$$

Where n is the number of objects.
The goal is minimize J_m and *Xie-Beni* functions simultaneously.

2.2 Solution representation and initial state

The type of representation of the approach is based on medoids. Each solution is coded as a vector of length k where it is kept as an identifier for each medoid, $k$being the number of clusters to identify. Starting from specification of number of clusters, are randomly selected k objects as medoids of the groups. Conformed initial state is conformed with this information. Each state has its own membership matrix, where all information related to belongingness of objects to clusters is stored.

2.3 Operator

Several operators are proposed,each of which leads to a different solutions. Due to medoid based representation, operators are designed to replace medoids for existing objects only.

2.3.1 Combination

Based on (Beausoleil 2007), selects all possible combinations from the vector of medoids and replaces them with objects randomly selected from data set. This operator is designed to create diversity.

2.3.2 Multiple flip

Based on the flip operator described in (Hruschka 2009), each medoid is substituted for objects from the data set. This operator makes more discrete changes because only it makes substitutions of length 1 in vector of medoids leading to stretch searching space.

2.3.3 Separator

It identifies the two medoids with the minor distance between them, let say, medoid A and B. It replaces B by each object whose dissimilarity with A is greater than A with B. Every possible substitution generates a different state. After, find all possible states keeping A and substituting B, it does the same procedure but keeping B and substituting A. This operator is intended to find solutions whose minimal separation Equation (19) is greater than the previous solutions.

2.3.4 Sequential strategy and Random strategy

These were extracted from (Dávila Ermus 2013). The strategy is switching the operator used in the procedure to generate neighborhood of state. In the first case there is a predefined order of switching whereas in the second the order of switching is random. As can be see, various operators are needed, thus previously described operators are used for switching strategies.

For further use and experimentation several of operators can be found in (Hruschka 2009) and (Martinez Pedroso 2014).

2.4 Procedure.

Let x_a current solution, x_c candidate solution, $V(x_a)$ neighborhood of current solution, L list of non dominated solutions, i.e. optimal Pareto set, x_{list} one solution from optimal Pareto set, U membership matrix

of a giving state and S the searching space. The term $o(x_a)$ indicates the application of a givenoperator to the current state in order to generate neighborhood $V(x_a)$.The procedure it executes is as follows

Algorithm 1. MultiObjective-HillClimbing for Partitional Clustering.

Take $x_a \in S$
Compute $U$of x_a with Equation (13).
Recomputemedoids in x_a with Equation (14).
Update $U$of $x_a$with Equation (13).
Add x_a to$L$
Repeat
  Apply $o(x_a)$ to generate $V(x_a)$
    Take $x_c \in V(x_a)$
    Compute $U$of x_c with Equation (13).
    Recomputemedoids in x_c with Equation (14).
    Update $U$of $x_c$with Equation (13).
    If $x_a$does not dominatex_c
    Repeat
    Take$x_{list} \in L$
    If $x_c$dominatesx_{list}
    Remove $x_{list}$from$L$
    EndIf
    Until end of list L or x_{list}dominatesx_c
    If $x_c$was not dominated
    Add$x_c$to$L$
    $x_a \coloneqq x_c$
    EndIf
EndIf
Until limit of iterations is reached

The procedure starts creating an initial solution x_a and calculating its fuzzy membership matrix U (steps 1 and 2). Then compute medoids, update U of current state x_a and add current state to$L$list (steps 3, 4 and 5). After this an iterative process starts from step 6 to 24, where in each iteration the operator is applied to the current solution (step 7), generating the neighborhood of the current solution. Stochastically is selected as a candidate solution from the neighborhood of current solution (step 8). Membership matrix U of candidate solution is computed with Equation (13), with this medoids of candidate solution ($\mathrm{x_c}$) are recalculated with Equation (14) in step 10, then the membership matrix of candidate solution is updated (step 11). Dominance verification between candidate solution and current solution is carried out (step 12). If candidate solution is non dominated by current solution, it triggers an iteration of L list (step 13 to 18), if a solution of L list is dominated by candidate solution, it is removed from the list. After the inner loop finishes, if candidate solution is non dominated by any of the solutions of L list, this is included in L and is taken as the current solution. The outer loop ends when it reaches the limit of iterations, stopping the procedure.

3. RESULTSAND DISCUSSION

3.1 External validity Indexes.

In order to measure the efficacy of proposed alternatives external validity indexes, Adjusted Rand Index (ARI) and Mincowski Score (MS) are used.

A clustering solution of n elements can be represented by a matrix of $n\ x\ n$ denoted as C, where $C_{ij} = 1$ if object i and object j are in the same cluster according to the known solution and $C_{ij} = 0$otherwise. If$T$is a matrix representing the correct clustering, let a, b, c, d respectively the number of pairs of points belonging to the same cluster in both T and C, the number of pairs belonging to the same cluster in T but to different clusters in C, the number of pairs belonging to different clusters in T but to the same cluster in C, and the number of pairs belonging to different clusters in both T and C. Adjusted Rand Index is defined in (Mukhopadhyay 2013) as:

$$ARI(T, C) = \frac{2(ad - b}{(a + b)(b + d) + (a + c)(c + d)} \qquad (21)$$

Where $0 \leq ARI(T, C) \leq 1$

An ARI value closer to 1 indicates a better solution with, $ARI(T, T) = 1$.

While giving the same matrixes T and C the Mincowski Score is defined as follows:

$$MS(T, C) = \frac{\|T - C\|}{\|T\|} \qquad (22)$$

where:

$$\|T\| = \sqrt{\sum_{i}^{n}\sum_{j}^{n} T_{i,j}} \qquad (23)$$

The Mincowski Score value is the normalized distance between two matrixes. The lower MS value, the better partitioning founded, with correct solution founded if $MS(T, C) = 0$.

3.2 Experiment strategy.

In the first phase all alternatives are tested in an application domain in order to contrast the performances in terms of partitioning quality. After this, the best alternativesare selected for comparison with relevant solutions presented in literature, in the same data set. Also, a comparison of mono-objective vs multiobjective optimization approaches is observed with classical mono objectives proposals, fuzzy k-means and fuzzy k-medoids, in numerical and categorical data domains respectively.

Two frequently used data sets named Iris and Zoo were selected to perform the experiments. The data sets were extracted from the UCI Machine Learning Repository available onhttps://archive.ics.uci.edu/ml/datasets.html. The next table describes data sets selected.

Table I.Data sets description.

Name	Objects	Numerical attributes	Categorical Attributes	Clusters
Iris	150	4	0	3
Zoo	101	0	16	7

Parameter settings. The maximum limit of iterations is fixed, up to 300. In case of the sequential strategy approach the operator will change after 50 iterations. The defined sequential order is first apply

Combination operator, then Multiple Flip operator and afterSeparator operator. With this tuning at least two loops of sequential strategy will be reached.

Table II. Validity indexes values of Hill Climbing on data sets "Iris" and "Zoo".

	Approaches	Jm	XB	ARI	MS
Data setIris	Separator operator	66.14	0.57	0.66	0.66
	Combinationoperator	62.30	0.22	0.80	0.51
	MultipleFlipoperator	62.15	0.23	0.79	0.52
	Sequentialstrategy	62.30	0.22	0.80	0.51
	Randomstrategy	62.30	0.22	0.80	0.51
Data setZoo	Separatoroperator	26.9432	0.1777	0.6268	0.7050
	Combinationoperator	25.5279	0.1289	0.8135	0.5189
	Multipleflipoperator	26.0422	0.1860	0.6707	0.6744
	Sequentialstrategy	25.7950	0.1349	0.8235	0.5024
	Randomstrategy	25.7782	0.1337	0.8266	0.5015

The Separator operator approach was designed with the objective of identifying clusters as far away as possible, however it can be observed that in both data sets Separator had the worst performance according to Xie-Beni index in comparison with all other approaches. Thus it can be said, according to the data sets tested, Separator operator does not accomplish its purpose. Nevertheless, Separator operator should be tested in different data sets.

On the other hand Combination operator, designed to generate diversity, shows the best results compared with other single operators (Combination and Multiple Flip) measured in ARI and MS values in both databases. This suggests that usinga higher diversity in the searching processcan obtain better results.
Further, switching strategies shows the best results in both databases according to external validity indexes values in contrast to single operators. This indicates switching strategies resulted in benefits from using various operators instead of just one, diversifying and stretching searching space every time the switch occurs, which allows for a more sophisticated exploration of the searching space.
Once approaches are compared,the best alternative proposed is Random strategy (based on previous results). It is then selected in order to contrast with other proposals in terms of ARI and MS values. Proposals considered in this phase optimize simultaneously J_m and *Xie-Beni*, and are able to model fuzzy context. One of the approaches is (MODEFCCD) which is proposed in (Saha 2013a), it was designed for categorical data, thus will be observed in data set Zoo only. On the other hand, the (MOMoDEFC) method developed in (Saha 2011) can only cover numerical domain, therefore data set Iris is suitable scope for comparison. As it was stated previously mono optimization techniques, Fuzzy k-means and Fuzzy k-medoids, were selected for Iris and Zoo respectively. ARI and MS values of Fuzzy k-medoids were extracted from (Mukhopadhyay 2013) and (Saha 2013a) respectively. In Fuzzy k-means case all values were obtained from (Saha 2011).

Table III.Validity indexes measures of Fuzzy k-means, MOMoDEFC and Hill Climbing with Random Strategy operator in numeric data set Iris.

	Approaches	J_m	XB	ARI	MS
Data set Iris	Fuzzy k-means	60.8520	0.3302	0.7832	0.4603
	MOMoDEFC	62.2102	0.1274	0.9342	0.2636
	Randomstrategy	62.30	0.22	0.80	0.51

Better optimization of theJ_m value suggests, in the Iris data set, the Fuzzy k-means approach identifies more compact groups than HC Random Strategy. However for XB values the opposite occurs, better results with HC Random strategy suggest clusters with larger separations between them. The approach developed in this research offers better ARI but worse MS values than the Fuzzy k-means approach, which suggests in this scenario, it cannot be determined precisely which method offers better results, nevertheless a starting point for further rigorous analysis is offered. MOMoDEFC indexes validity values are worse but similar with the exception of the MS value, where there is a major difference.

Table IV.Validity indexes measures of Fuzzy k-medoids, MODEFCCD and Hill Climbing with Random Strategy operator in numeric data set Zoo.

	Approaches	J_m	XB	ARI	MS
Data set Zoo	Fuzzy k-medoids	-	-	0.7121	0.4313
	MODEFCCD	-	-	-	0.2461
	Random strategy	25.7782	0.1337	0.8266	0.5015

In the Zoo data set, the Fuzzy k-medoidsapproach offers better results for the MS value but worse ARI values with respect to the HC Random strategy approach. It can be outlined that the same phenomenon of contrasting results occurs in both data sets when mono objective optimization techniques and multiobjective HC Random Strategy approaches are compared according to ARI and MS values.
On the other hand, MODEFCCD provides better results with respect to MS over the HC Random Strategy approach.

4. CONCLUSIONS

Hill climbing needed to be adapted from its original definition in order to apply Pareto optimal principles. Although, this has been done in the past, according to literature surveyed, hill climbing had not been used to face multiobjective partitional clustering, therefore a novel alternative to tackling such a problem is present here. Moreover, the solution presented is capable of covering fuzzy domain and mixed data.

Multiple operators were designed in order to explore the neighborhood of each state taking into account themedoid representation previously adopted.

After experiments were carried out, switching strategies show the best results in both databases according to external validity indexes values in contrast to single operators. This indicates switching strategies resulted in benefits from using various operators instead of just one, diversifying and stretching searching space every time that a switch occurs, which allows a more sophisticated exploration of searching space.

5. FUTURE WORK

More experiments need to be done in order to run significant test and obtain stronger evidence comparing different alternative presented in this research against solutions surveyed in literature.Due to necessity of search the space in an efficient manner and taking into account that switching strategies show better results, more operators can be designed in order to feed this approach. Also the possibility of using prototype representation offers the opportunity to design different and specific operators.

Different alternatives presented in this research have been only applied to experimental data set. Therefore, these can be applied to real data sets and measure it efficacy.More algorithms can be developed based on

local search meta heuristics (such as Hill Climbing) in order to tackle complex problem of performing multiobjective partitional clustering.

6. ACKNOWLEDGMENTS

The author expresses his most sincere gratitude to close relatives and friends who supported him and encouraged him in hispursuit of this investigation.

REFERENCES

[1] ANIRBAN MUKHOPADHYAY, U. M. Multiobjective Approach to Categorical Data Clustering. In *IEEECongressonEvolutionaryComputation(CEC 2007.* IEEE, 2007, p. 8.

[2] BANDYOPADHYAY, S., SRIPARNA SAHA, UJJWAL MAULIK, KALYANMOY DEB; A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 2008, 12(3), 15.

[3] BANDYOPADHYAY, S., UJJWAL MAULIK, ANIRBAN MUKHOPADHYAY Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery. IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 2007, 45(5), 6.

[4] BEAUSOLEIL, R. P. Multiobjective Clustering using Tabu Search 2007.

[5] BENI, G., XUANLI LISA XIE A validity measure for fuzzy clustering. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 8 de Agosto 1991, 13(8), 7.

[6] CABALLERO, R., MANUEL LAGUNA, RAFAEL MARTÍ, JULIÁN MOLINA Multiobjective Clustering withMetaheuristic Optimization Technology 2008.

[7] CABALLERO, R., MANUEL LAGUNA, RAFAEL MARTÍ, JULIÁN MOLINA Scatter Tabu Search for Multiobjective Clustering Problems 2009.

[8] COELLO COELLO, C. A., GARY B. LAMONT, DAVID A. VAN VELDHUIZEN *Evolutionary algorithms for solving multiobjective problems*. edited by J.R.K. DAVID E. GOLDBERG. Edtion ed.: Springer, 2007. 810 p. ISBN 978-0-387-33254-3.

[9] CHUN-WEI, T., WEN-LING CHEN, AND MING-CHAO CHIANG. A Modified Multiobjective EA-based Clustering Algorithm with Automatic Determination of the Number of Clusters. In *International Conference on Systems, Man, and Cybernetics.* COEX, Seoul, Korea: IEEE, 2012, p. 6.

[10] DÁVILA ERMUS, L. Propuesta para la Construcción Automática del Horario Docente de la Facultad de Ingeniería Informática del Instituto Superior Politécnico "José Antonio Echeverría" aplicando algoritmos metaheurísticos. Instituto Superior Politécnico "José Antonio Echeverría", 2013.

[11] DI NUOVO, A. G., MAURIZIO PALESI, VINCENZO CATANIA. Multi-Objective Evolutionary Fuzzy Clustering for High-Dimensional Problems. In. Catania, Italy: Universita di Catania, 2007, p. 6.

[12] DÍAZ PANDO, H., SERGIO CUENCA ASENSI, ROBERTO SEPÚLVEDA LIMA, ALEJANDRO AND J. F. C. ROSETE SUÁREZ Algoritmos metaheurísticos en el problema del particionado hardware/software de sistemas embebidos. INTELIGENCIA ARTIFICIAL, 2013, 16(51), 14.

[13] DÍAZ, R., ALEJANDRO ROSETE SUÁREZ. A STUDY OF THE CAPACITY OF THE STOCHASTIC HILL CLIMBING TO SOLVE MULTI-OBJECTIVE PROBLEMS. In. Ciudad Habana, Cuba: Instituto Superior Politécnico José Antonio Echeverría 2001, p. 4.

[14] DUTTA, D., PARAMARTHA DUTTA, JAYA SIL 2012a. Clustering by Multi Objective Genetic Algorithm. In *Proceedings of the 1st Int'l Conf. on Recent Advances in Information Technology*, RAIT, India2012a IEEE.

[15] DUTTA, D., PARAMARTHA DUTTA, JAYA SIL. Clustering Data Set with Categorical Feature Using Multi Objective Genetic Algorithm. In *International Conference on Data Science & Engineering (ICDSE).* 2012b, p. 6.

[16] DUTTA, D., PARAMARTHA DUTTA, JAYA SIL. Data Clustering with Mixed Features by MultiObjective Genetic Algorithm. In *12th International Conference onHybrid Intelligent Systems (HIS.* IEEE, 2012c, p. 6.

[17] DUTTA, D., PARAMARTHA DUTTA, JAYA SIL. Simultaneous Feature Selection and Clustering for Categorical Features Using Multi Objective Genetic Algorithm. In *12th International Conference onHybrid Intelligent Systems (HIS).* IEEE, 2012d, p. 6.

[18] DUTTA, D., PARAMARTHA DUTTA, JAYA SIL. Simultaneous Continuous Feature Selection and K Clustering by Multi Objective Genetic Algorithm. In *3rd IEEE International Advance Computing Conference (IACC).* IEEE, 2013, p. 6.

[19] HANDL J., J. K. Evolutionary Multiobjective Clustering 2004a.

[20] HANDL J., J. K. Multiobjective clustering with automatic determination of the number of clusters. Manchester: 2004b.

[21] HANDL J., J. K. Multiobjective clustering around medoids 2005, 8.

[22] HANDL J., J. K. An Evolutionary Approach to Multiobjective Clustering. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 2007, 11(1), 21.

[23] HRUSCHKA, E. R., RICARDO J. G. B. CAMPELLO, ALEX A. FREITAS, ANDRÉ C. P. L. F. DE CARVALHO A Survey of Evolutionary Algorithms for Clustering. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2009, 39(2), 22.

[24] HUANG, Z. Extensions to thek-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 1998, 2, 22.

[25] INFANTE ABREU, A. L., DRA. MARGARITA ANDRÉ AMPUERO, DR. ALEJANDRO ROSETE SUÁREZ, LALCHANDRA RAMPERSAUD Conformación de equipos de proyectos de software aplicando algoritmos metaheurísticos de trayectoria multiobjetivo. INTELIGENCIA ARTIFICIAL, 2014, 17(54), 16.

[26] KAMBER, J. H. Y. M. *Data mining: Concepts and Techniques*. edited by M.R. JIM GRAY. Edtion ed. San Francisco: Morgan Kaufmann Publishers, 2006. 772 p. ISBN 13: 978-1-55860-901-3.

[27] MARTINEZ PEDROSO, P. B., DARIAN HORACIO GRASS BOADA, ISABEL MARIA HIGUERA IGARZA (2014). Extensión de la herramienta BiCIAM a partir de un algoritmo de agrupamiento multiobjetivo utilizando metaheurística de trayectoria simple. Facultad 2. La Habana, Cuba, Universidad de las Ciencias Informáticas.78.

[28] MATAKE, N., TOMOYUKI HIROYASU, MITSUNORI MIKI, TOMOHARU SENDA Multiobjective Clustering with Automatic k-determination for Large-scale Data 2007, 8.

[29] MUKHOPADHYAY, A., UJJWAL MAULIK. Multiobjective Approach to Categorical Data Clustering. In *IEEECongressonEvolutionaryComputation(CEC 2007.* IEEE, 2007a, p. 8.

[30] MUKHOPADHYAY, A., UJJWAL MAULIK, SANGHAMITRA BANDYOPADHYAY. Multiobjective Genetic Fuzzy Clustering of Categorical Attributes. In *10th International Conference on Information Technology.* IEEE, 2007b, p. 6.

[31] MUKHOPADHYAY, A., UJJWAL MAULIK, SANGHAMITRA BANDYOPADHYAY Multiobjective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 2009, 13(5), 15.

[32] MUKHOPADHYAY, A., UJJWAL MAULIK, SANGHAMITRA BANDYOPADHYAY. Simultaneous Informative Gene Selection and Clustering through Multiobjective Optimization. In., 2010, p. 8.

[33] MUKHOPADHYAY, A., UJJWAL MAULIK, SANGHAMITRA BANDYOPADHYAY. Hybrid Evolutionary Multiobjective Fuzzy C-Medoids Clustering of Categorical Data. In *IEEE Workshopon HybridIntelligentModelsand Applications (HIMA).* IEEE, 2013, p. 6.

[34] SAHA, I., UJJWAL MAULIK, DARIUSZ PLEWCZYNSKI A new multi-objective technique for differential fuzzy clustering. Applied Soft Computing, 2011, 11, 12.

[35] SAHA, I., UJJWAL MAULIK, DEBASREE MITY. Categorical Data Analysis using Multiobjective Differential Evolution based Fuzzy Clustering. In *2013 InternationalConference onAdvances inComputing,CommunicationsandInformatics (ICACCI.* IEEE, 2013a, p. 5.

[36] SAHA, S., SANGHAMITRA BANDYOPADHYAY A generalized automatic clustering algorithm in a multiobjective framework. Applied Soft Computing, 2013b, 13, 20.

[37] SAHA, S., SANGHAMITRA BANDYOPADHYAY A new multiobjective clustering technique based on the concepts of stability and symmetry. Knowl Inf Syst, 2009, 23(1), 27.

[38] SMITH, K. I., RICHARD M. EVERSON, JONATHAN E. FIELDSEND, AND C. M. A. R. MISRA Dominance-Based Multi-Objective Simulated Annealing. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 2005.

[39] SURESH, K., DEBARATI KUNDU, SAYAN GHOSH, SWAGATAM DAS, AJITH ABRAHAM Data Clustering Using Multi-objective Differential Evolution

[40] Algorithms. Fundamenta Informaticae, 2009, 97, 25.

[41] XU, R., DONALD C. WUNSCH II *CLUSTERING*. edited by D.B. FOGEL. Edtion ed. New Jersey, USA: John Wiley & Sons, inc., 2009. 364 p. ISBN 978-0-470-27680-8.