Ca' Foscari University

Department of Computer Science

MASTER THESIS

Game Theoretic Approach to Multiple Similarity

Venice, March 1, 2013

By:

Eyasu Zemene

Supervisor:

Prof. Marcello Pelillo

Acknowledgments

I would like to thank my supervisor,Prof. Marcello Pelillo for his great support for my thesis.His way of giving explanations and the way that he gave some motivations helped me understand and do my work easily. I would also like to thank Dr. Samuel Rota for his help giving me solutions for some problems, and Prof. Andrea Torsello giving me some advice and work on papers related to my thesis. Finally I would like to thank Prof. Salvatore Orlando for his help in solving some personal problems so that I can do my work freely.

ABSTRACT

Most of the techniques from fields of computer science are now being used by most researchers to solve a whole range of the real world problems in Life science, vision and graphics, networking and artificial intelligence, performance and reliability, distributed systems etc... . Some of the important techniques include techniques from AI and Computer Vision. Most of the techniques modeled so far to solve the real world problems, however, have limitations and also used many assumptions.As an example if we see the clustering problem, most of its domain is dominated by the partitional approach which has many limitations and uses many assumptions as symmetricity of similarity, non overlapping of clusters ...etc. There are approaches (which we used to develop different models) that avoid most of the limitations and also perform better.

The aim of my thesis is to develop a model based on the notion of game theory and novel graph-theoretic approach, dominant set,which has been shown to be important in many fields as: Image segmentation [**?**], Bioinformatics: [**?**], [**?**],Security and Video surveillance: [**?**], [**?**], Content-based image retrieval: [**?**], Video analysis, object tracking,human action recognition, Multiple instance learning, Feature selection [**?**], Image Matching and Registration [**?**] etc .... All the above applications used the conventional game theoretic approach which can't recognize (in the case of multiobjective) relations among the different similarity measures. It is known that there are many practical problems in the real world that are not with a single objective, so in this case it is better to think about multi-objectives than the conventional game theoretic problems.It means that there are many practical scenarios where single objective case become handicapped in recognizing the different relations among various measures,the similarities,that players perceive in an evolutionary sense [M.] . So we can say in such situations, models

using notions of multiobjective game theoretic approach outperforms models from single objective game theoretic approach.

In multicriteria games, evolutionary stability has been given little attention. Recently,however, a formal definition of an evolutionary stability for general vector orders was proposed by Somasundaram and Baras [M.].They also show that equivalence of the player's best strategy against opponents to a parametric multi-objective linear program and this relation is useful in constructing the trade-off weights among the various payoffs,objective similarities. We used these weights to combine the different similarity measure as these days applications show that using different similarity measures combined instead of a single similarity measures is effective.Since it is very difficult to know the exact similarity measure for the data it is better to combine different similarity measures and use it as a representation of the data.We demonstrate the potential of this work doing experiments with various similarity matrices built from different datasets from UCI and toy datasets generated by ourselves , and the results confirm the proposed approach.So models using this approach (with dominant set) have many importance: They outperform models (in multiobjective cases) where each partition is given an equal weight in the combination process and all similarities in each partition contribute to the combined solution equally, they avoid most of the limitations of other models as they use the game theoretic approach. We can say in general the proposed approach is an efficient one that can solve many real world problems from multidisciplinary environments

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clustering is an unsupervised learning which learns by observation rather than using the labels.Different fields like machine learning,bioinformatics, information retrieval, pattern recognition, image analysis and others have been using it for different analysis for the application of practical problems. There are a number of clustering algorithms that differ in their notion of what its constituent is and the techniques used by them to find them in an efficient way.Partitional, hierarchical, Spectral, density based and grid based are well known and mostly used clustering algorithms. All of them use different technique to attack the problem and so all have their own advantages and disadvantages. Even-though there are plenty of such algorithms, till now its notion is not defined very well. But all of them have same goal which is to increase the internal homogeneity and the non-homogeneity with the external. From those algorithms listed above the well known, mostly used and an interesting one is the partitional paradigms because of its oversimplified formulation where powerful ideas are employed from the technical tools as which can be linear algebra, graph theory, optimization theory, information theory, etc. Even-though this has such an attractive formulation it has many limitations. The well known limitation of the partitional paradigm is that the number of clusters to which all the objects are assigned must be known. But even knowing this parameter there is a problem of deciding what a cluster constitutes for example we may end up in a completely wrong cluster result base on the initial centroid chosen for each cluster. To overcome this problem (which is another additional task) different approaches are used

as genetic algorithm, simulated annealing,harmony search techniques and ant colony optimization.

At the very beginning of this thesis we are going to discuss about the partitional paradigm and its limitations and in the up coming sections we present the motivation of our approach and the goal in which some background on game theoretic notion of a cluster is covered.In the last part of this chapter, presented is the organization of the rest of the thesis.

1.1 The Partitional clustering approach and limitations

It is known that the field of clustering is dominated by the partitional algorithms as the majority approach to clustering problem is to partition the set of input data and even sometimes they said it is defined as the problem of partitioning data.

The first limitation of the partitional paradigm is that each object should be assigned to one of the clusters. But there are many applications where this is not a useful property. For example if we see the area of computer vision there is a problem which we commenly face that tries to separate the figure from the base (the ground) in which some structures pops out from the ground and and there are some clutter (noise parts).We can see the following examples for this limitation.

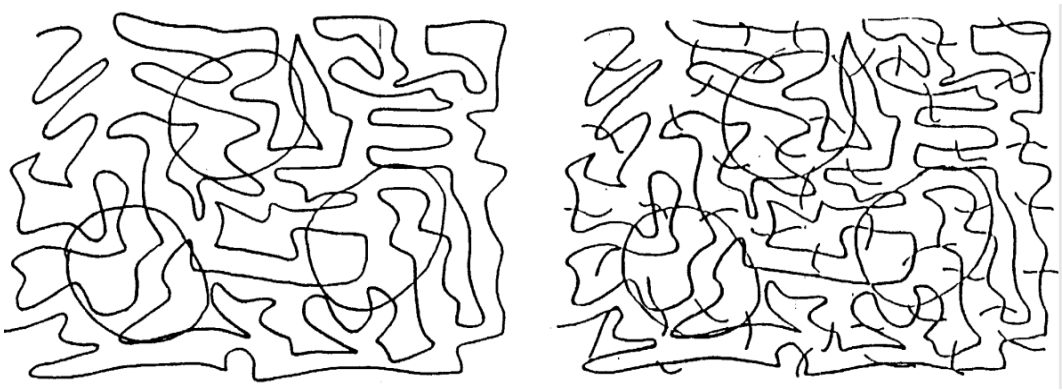


Fig. 1.1: *the first figure the prominent blobs are perceived immediately and with little effort. Locally, the blobs are similar to the background contours (adopted from Mahoney (1986)), in the second part intersections were added to illustrate that the blobs are not distinguished by virtue of their intersections with the background curves*

Here as we can see there is a clutter part that we don't want to include in a cluster.

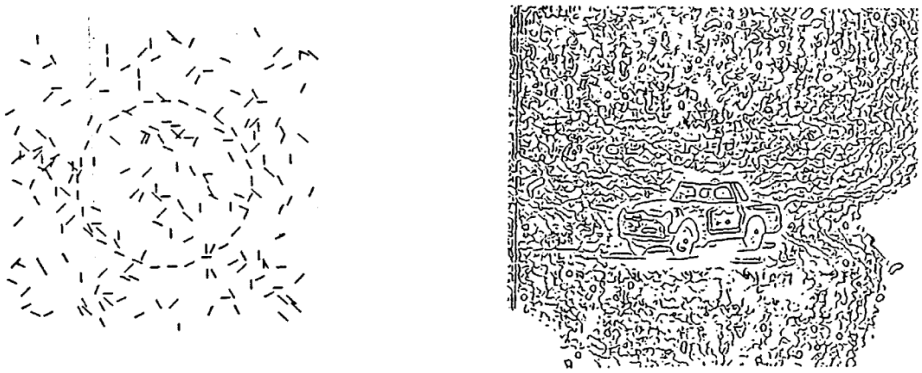


Fig. 1.2: *A circle in a background of 200 randomly placed and oriented segments. The circle is still perceived immediatly although its contour is fragmentd. The second is an image of a car in a cluttered background. Their attention is drawn immediatly to the region of interest. Itseems that the car need not be recognized to attract their attention. The car also remain salient when parallel lines and small blobs are removed, and when the less textured region surrounding parts of the car is filled in with more texture.*

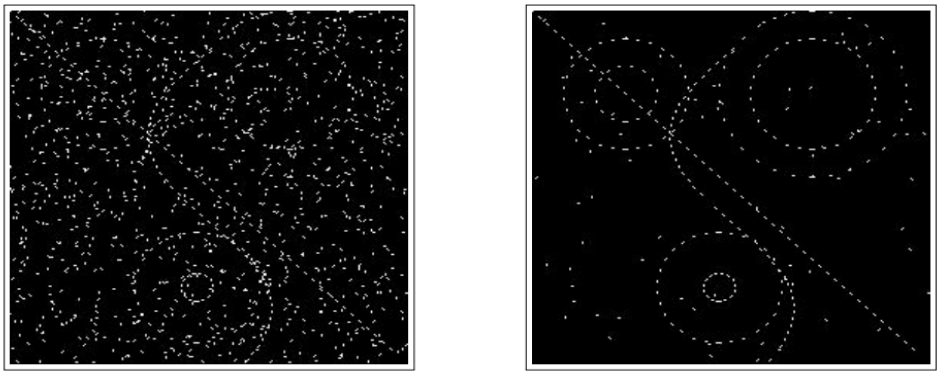


Fig. 1.3: *On the first part is a synthetic image with 1250 elements. Circle, straight line and sinusoids are plugned into a randomly generated element. The second figure is the result of applying mean field annealing to the syntehtic image(Adopted from: Herault, L., Horaud, R)*

Here is another example that shows the limitation and also show the performance of the game theoretic notions of a cluster.

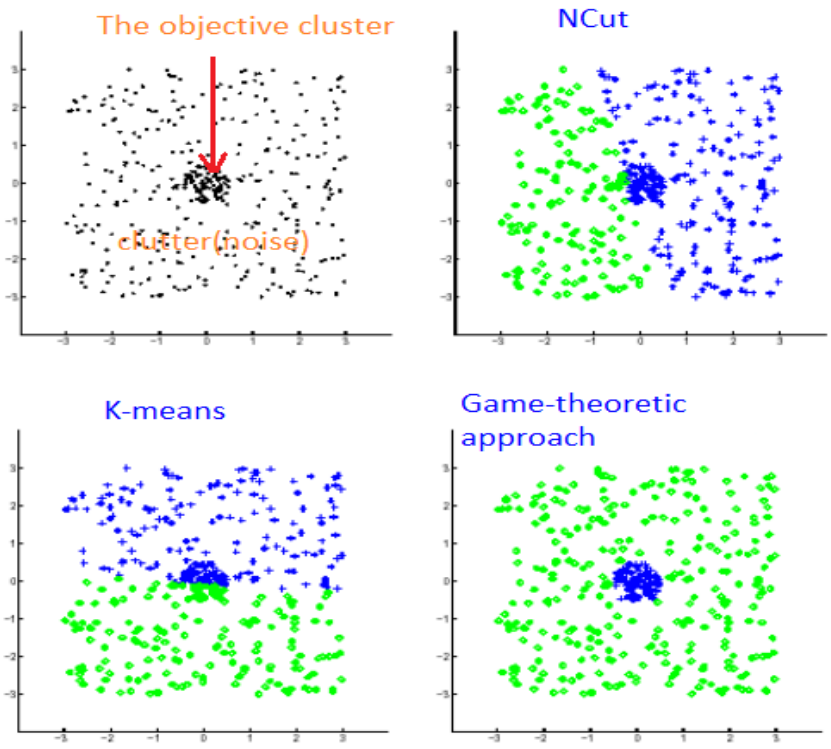


Fig. 1.4: *A structure with a clutter and the cluster resulted from different approaches*

As we can see from the resulted clusters, the Normalized cut and the K-means approach can't give the required structure. As they have to assign all the objects to one of the two clusters, they also clustered the clutter which we don't want to be included in any of the clusters.

The second limitation of the partitional approach is the need of overlapping clusters.The partitional approach impose that each element can't belong to more than one cluster. But there are applications where this requirement is needed very badly.In this paper

[?] the authors pointing majority of clustering algorithms are partitional,they inform many real world datasets have inherently overlapping clusters. As an example they point out that recent biological datasets are frequently overlapping and has led to new clustering models that allow hard assignment of data points to multiple clusters. K.A. Heller and Z. Ghahramani wrote in their paper [?] about the gene that may have many functions,therefore belonging to several distinct clusters of genes, and also they point out that a biologist may want to discover these through unsupervised modelling of gene expression data. Some other applications include: clustering documents into topic categories, perceptual grouping and segmentation of image with transparent surfaces.Here is a simple example that shows the case when there is a group overlap.

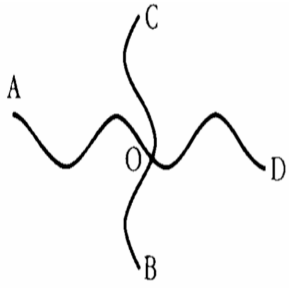


Fig. 1.5: *To which structure does O belongs to, AD or BC ?*

Torsello, Samuel and Pelillo showed in their paper [?] that dominant set clustering approach outperform both the Normalized cut (Partition) and minimal shift embedding (Partition) approach. The figure below is one of the results the used to show how it outperforms.

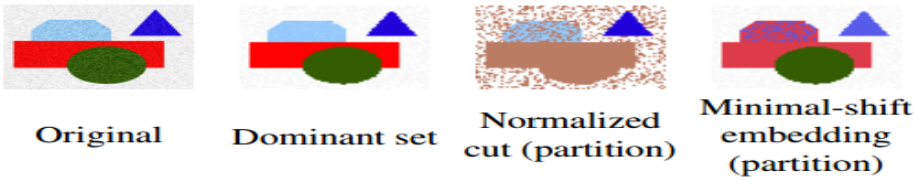


Fig. 1.6: *Result of Synthetic segmentation done by Torsello, Samuel & Pelillo*

The other limitation is not the direct limitation of the partitional approach, instead

it arises from the technical tools typically used to attack the problem. In the partitional paradigm, all the technical tools it can be linear algebra,graph theory, Information theory and others assumes the similarities between the data to be clustered are symmetric (and non-negative). But Amos Tversky gave in his paper [**?**] an empirical evidence for asymmetric similarities and argues that similarities should not be treated as a symmetric relation. The asymmetric similarity also come up in many applications as: Kullback-Leibler divergence of the probabilistic measure, normalized edit distances, in the coalition of images shapes or sequences in a template matching process. In addition to this, David,Daphna and Yoram used in their paper [**?**] a non metric distances for image retrieval and class representation.

1.2 Suggested Solution for the limitations

pelillo and Samuel, pointing probabilistic model-based approaches do not suffer from most of the limitations discussed here,they propose another type of strategy which doesn't depend on the partitional process, instead they try to focus on the very notion of the cluster,how they can formalize it and drive algorithm for finding the clusters. By doing so, they try to answer the question "What is a cluster?" instead of the question "how can we partition the data?". Because they taught that finding an answer to thier question clarify the very nature of clustering problem and overcome the major limitations and deal with more jeneral problems which means the clusters may overlap, noises not be assigned and similarities may be symmetric and also they can be negative. What they used to answert this question is that of the notion of evolutionary game theory and its equilibrium. They formulate the problem as a non cooperative game and find the evolutionary stable strategies using the discrte repilicator dynamics.The result that they we get in this way satisfies both the internal and external criterion of a cluster

(which informally defined as a maximally coherent class), and so both concepts of a cluster and an equilibrium from evolutionary game theory coincides.This game theoretic notion of a cluster has wide applications in many fields as machine learning,computer vision,bioinformatics, ... etc. It has also many intersiting features like: No need to know the number of clusters, no assumption on data representation and the structure of the affinity matrix, clutter elements are not included in any of the clusters,allows extraction of overlapped clusters ,and it generalizes naturally to hypergraph clustering problems. In the next section we are going to discuss about the thesis which is based on this game theoretic clustering approach and also about the organization of the rest of the thesis.

1.3 Motivation and our goal

Starting from the past many researchers have been trying to solve problems like multiple similarities. It is very important to transform or represent the data in a way we can solve the problem easily.When we map our data to an appropriate feature space, it is very important to choose the right similarity measure that represents our data as it has a great impact on the performance of an algorithm that solves the problem.Thinking that combining different similarity measures in to one to represent the data help solve problems effectively and also since it is very difficult to know the best representation of our data,it become common to create different representations for a data and combine in different way to use it as a representation of the data. Whether it is combined or not almost all the works that are done to solve such problems are specific in the sense that they use a very specific similarity measures. For examples in the papers [KZ03] , [HCHC11] used only a positive semi-definite matrices. Here in our case we are working in a generalized version.To make this representation general we used the game theoretic approach discussed in the previous section.This work which is based on the new graph-

theoretic concept that is the 'dominant set' can be applied on all previous works that uses the notion of dominant set. Since this work is based on the multiobjective games nothing can prevent it from giving us a better result.

The rest of the thesis is organized as follows: In the next chapter we present the concept of game theory, Nash equilibrium, and evolutionary game theory. Chapter three is about the very notion of a cluster based on dominant set and its connection to an Evolutionary Stable Strategies. Our proposed approach will be presented ,with some experimental results from toy and real dataset, in chapter 4. Finally we will discuss about some applications and future works and conclude the work.

Chapter 2

Game Theory

2.1 Introduction to Game Theory

Game theory was introduced first,mathematically,by Von Neumann and Oskar Morgenstern in the 1940.Their mathematical theory was restricted to the zero-sum-games in which the gains and the looses are equal. John F. Nash break this restriction and showed the very clear distinction between cooperative and non cooperative games which helps to use the importance of game theory in the real world.When the sum of the payoffs are not equal to zero or when the sum is no longer constant,maximizing and minimizing of the player's payoffs are no longer equal.Two players general sum game can be given as a matrix of elements with each elements have a pair of numbers that represents payoff of player I and player II respectively, or we can represent it as two matrices one for each player's payoff. For example if you see the following matrix player I is the row player and has 2 strategies while player II is the column player and has 3 strategies.We can represent the sets of the payoffs as two sets R and C so that the cross product ,RXC, represents the whole set(the whole elements of the matrix). If player I choose $r \in R$ and player II choose $c \in C$ then player I will get the payoff $u_1(x, y)$ and player II will get the payoff $u_2(x, y)$. Where $u_1(x, y) and u_2(x, y)$ are the payoff functions. 3×3 matrix

$$\begin{pmatrix} (1,2) & (1,-2) & (3,4) \\ (-1,-2) & (-2,1) & (0,-1) \end{pmatrix}$$

Here if player I choose r=1 and player II choose c=2, player 1 will get the payoff of $u_1(1,2) = 1$ and player II get the payoff $u_2(1,2) = -2$.

Let's see the two varieties of strategies: **Pure and Mixed Strategies**: When we see pure strategies a game is described by *the set of players I, the set of strategies S and the payoff functions U*.For this case we can see as an example a well known game,**Prisoners Dilemma:** The payoff matrices in different forms are as follows

$$\begin{array}{ccc} \text{Strategies} & defect & cooperate \\ defect & (-5,-5) & (0,-9) \\ cooperate & (-9,0) & (-1,-1) \end{array}$$

Which can be written as two matrices like this

$$PlayerI'spayoff = \begin{pmatrix} -5 & 0 \\ -9 & -1 \end{pmatrix}, PlayerII'spayoff = \begin{pmatrix} -5 & -9 \\ 0 & -1 \end{pmatrix}$$

As we can see from the matrix, both players will be better off is they choose the first strategy which is defect than the first one,so a player choose their dominant strategy irrespective of the strategy chosen by the other player.That means a player choose his best response.This is because both the first row and the first column dominates the second row and column respectively.Choosing the dominant strategy both players will get a penalty of 5 where as if both play the dominated strategies they will get a penalty of 1. Since the players will be better off when they both choose the dominant strategy, the game is called prisoners dilemma.It is called like this because of the following, two criminals one with big and one with small crime have captured and the police take them in to two different rooms. The police don't know which one is the bigger criminal, so he tell to them If one confesses and the other deny then the one who deny will be sent to

jail for 9 years while the other will be free.If both confess,they will be in jail for 5 year otherwise If both deny both will be in jail for 1 year.

To solve the above problem,we can make a circle on those payoff values chosen by the players and those which are circled pairs are the solution to the game. Here is the solution for the above game.

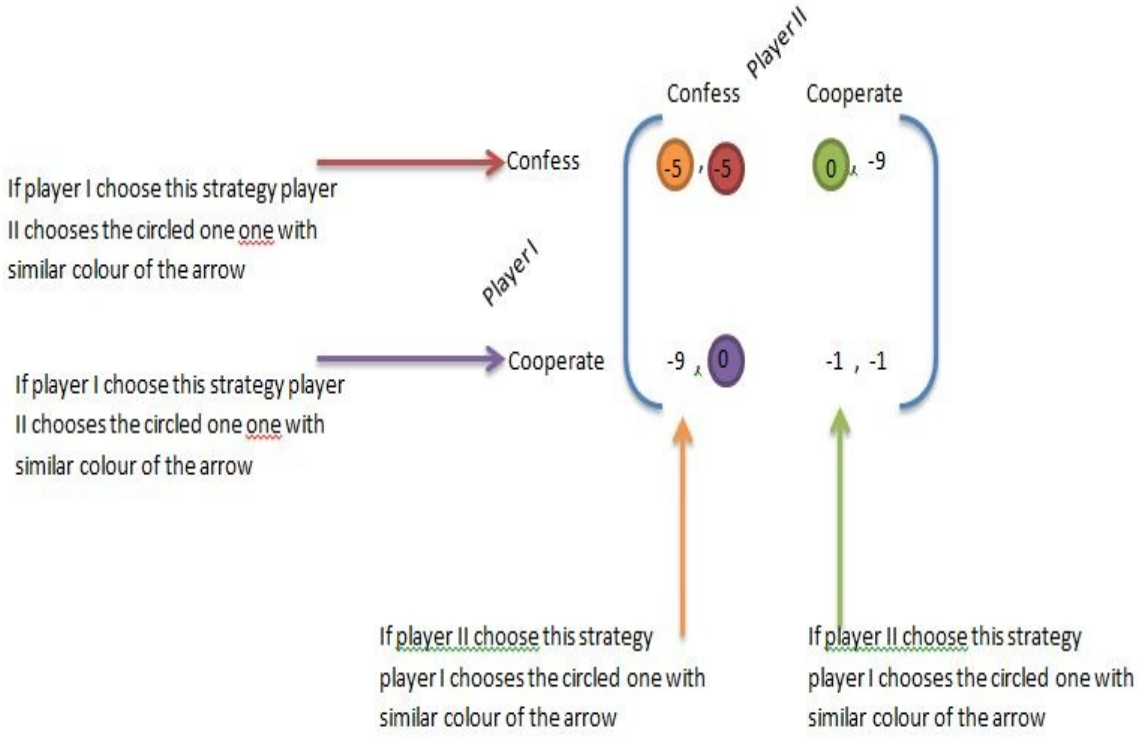


Fig. 2.1: Prisoners Dilemma

Here is the equilibrium point that the two players reached.

As we see in the above example we get an equilibrium point,the Nash equilibrium. Can we get this equilibrium point for all games? In pure strategies we can't where as in mixed strategies we can have the equilibrium point for all games.Let's see what mixed strategies means and see an example.

Mixed Strategies: The notion of mixed strategy is simply taking the probability distribution over the pure strategies of a players.If x_i is the mixed strategy for player

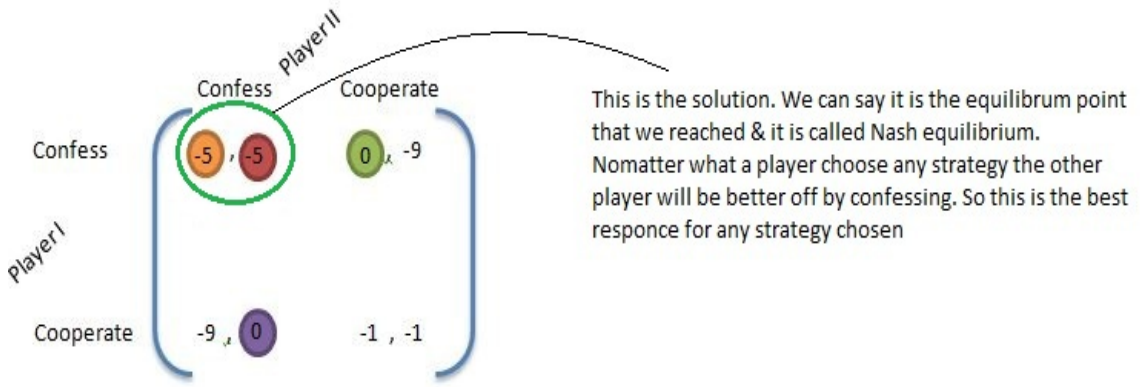


Fig. 2.2: The equilibrium point

i it is represented as a vector in R^m dimentional space. We know the probablities are positive and the summed up to one, so it is in the standard simplex. i.e $x_i \in \Delta_i$ and $\forall h \in \{1....m\}\ if x_{ih} > 0$ it is called the support of the mixed strategy which is $\sigma(x_i)$. For each players i in a set of players I,We have a mixed strategy x_i in the standard simplex Δ_i, a set that contains the whole set of this mixed strategies for all the players $x = \{x_1....x_{|I|}\}$ is called **mixed strategy profile**. The Cartesian product of the simplexes, $\Theta = \Delta_1 \times \Delta_2....\Delta_{|I|}$ gives us the **mixed strategy space**.If we have a unit vector $e_i^h = (0, 0, 0, 1, 0, ..., 0)$ in the m-space, it represents the h^{th}vertex of the simplex Δ_i which is the h^{th} pure strategy.If we have payoff matrices of A_i for all players $i \in I$, the payoff that player i get when he played against player j is $U(x_i) = x_i^T A_i x_j$.If x_i is a mixed strategy which gives a higher payoff ,than any other mixed strategies give, for player i against the strategy x_j then we say x_i is the best response for player i against the mixed strategies x_j.We can define it simply as follows.

Definition: Best Reply A strategy profile x_i^* is a best response to the strategy x_{-i} of the other players if

$$U_i(x_i^*, x_{-i}) \geq U_i(x_i, x_{-i}) \ \forall x_i \in \Delta_i$$

It solves the following maximization problem $\max_{x_i} U_i(x_i, x_{-i})$

The best reply in the extreme case is, in pure strategy, unique, but usually the best replies are always infinite and also if the best reply includes two or more pure strategies, any mixture of these strategies must also be a best reply. Or we can say If a mixed strategy is a best response then each of the pure strategies involved in the mix must itself be a best response. In particular, each must yield the same expected payoff.

Nash Equilibrium

Definition: A strategy profile $(x_1, x_2, x_3, ..., x_{|I|})$ is a Nash equilibrium if for all i $x_i \in$ best response of x_{-i} (for each player, his choice x_i^* is the best response to the other player's choice x_i^*). That means it is the best response to it self. We can write it in this form $x^T Ax \geq y^T Ax$ for all mixed strategies y.

$U_i(x_i, x_{-i}) \geq U_i(x_i^*, x_{-i}) \ \forall x_i^* \in \Delta_i$, and we get **Strict Nash Equilibrium** with strict inequality $\forall x_i^* \neq x_i$

Let's see the well known example which is Rock, Scissor,Paper game:

$$|x| = \begin{cases} x & \text{if } x \geq 0; \\ -x & \text{if } x < 0. \end{cases}$$

Player I

	Game	Rock	Scissor	Paper
	`Rock`	(0,0)	(1,-1)	(-1,1)
player II	`Scissor`	(-1,1)	(0,0)	(1,-1)
	`Paper`	(1,-1)	(-1,1)	(0,0)

Let's try to solve the problem in pure strategy i.e the case in which only one action is played at a time

Player I

Player II		Rock	Scissor	Paper
	Rock	(0 , 0)	(1 , -1)	(-1 , 1)
	Scissor	(-1 , 1)	(0 , 0)	(1 , -1)
	Paper	(1 , - 1)	(-1 , 1)	(0 , 0)

Player I

Player II		Rock	Scissor	Paper
	Rock	(0 , 0)	(1, -1)	(-1 ,1)
	Scissor	(-1 ,1)	(0 , 0)	(1, -1)
	Paper	(1, - 1)	(-1 ,1)	(0 , 0)

As we can see from the selected points, we have no any indices that are with pair circled (chosen) elements. So we can say we have no any Nash equilibrium in this pure strategy. But we know from Nash theorem every finite game has a mixed strategy Nash equilibrium. Let's try to solve the mixed strategy Nash equilibrium.

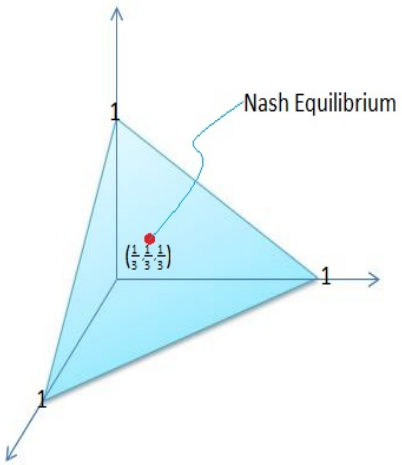
To solve the problem in this case, let's assume player I assigns the probabilities $p_r, p_s, and (1-p_r-p_s)$ for the rock, scissor and paper respectively. The for player II,

Row, Rock payoff aganist $(p_r, p_s, and (1-p_r-p_s)) = p_r*0+p_s*1+(1-p_r-p_s)*-1$

Row, Scissor payoff aganist $(p_r, p_s, and(1-p_r-p_s)) = p_r*-1+p_s*0+(1-p_r-p_s)*1$

Row, Paper payoff aganist $(p_r, p_s, and(1-p_r-p_s)) = p_r*1+p_s*-1+(1-p_r-p_s)*0$

If we solve all these payoffs to be equal we will get $\frac{1}{3}$ for every probabilities assigned. So our Nash equilibrium will be $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$



Why people study *games*?

Most of the real world situations in science and engineering can be modelled using games. Due to this we have seen most of the application of *game theory* in the fields of economics, biology,networking in telecommunication[EARJTAEL06],in network routing, in resource allocation in distributed systems,resource allocation in information and service transaction in internet[KP97].When we come to our case, so far we where describing the framework performing clustering assuming symmetry and non-negativity of the affinity matrix. When we come to the most general case, there are situations that you want a non symmetric or affinity matrix for example the most important distance measure in machine learning and statistics,Kullback-Liebler(KL) divergence for a distance measure of probability distributions, is asymmetric. There are other many cases in which the affinity matrix is asymmetric and negative[JS] [MM]. The standard approach in this

case is to symmetrize the affinity matrix [MM], but symmetrizing the matrix make it loose some of the important informations that are in the asymmetry. So which method is better for this case? We know that we no longer have a single objective function that can capture the structure of the data, so we need to rely on something more sophisticated and that is game theory. Game theory breaks the above limitations of the single objective optimization problem. Here is our point to answer the question how can one define the notion of a cluster when the affinity matrix are non symmetric and or negative.

Now a days, in the machine learning community peoples use the algorithmic aspect of game theory to design an algorithm for computational purpose. The notion of the algorithmic game theory was started by Nisan and Ronen [NR99]. In 2002, Ortiz and Kearns uses game theory for computational purpose in the study of machine learning. Here in our clustering case we use game theory for modelling a problem not for computational purpose.

2.2 Evolutionary Game Theory

Evolutionary game theory was introduced in the 18^{th} century by British theoretical evolutionary biologist and geneticist John Maynard Smith who studied evolution and has related animal behaviour to game theory and fighting strategies.His three important theories: Game theory, Evolutionary stable strategies,and hawk and dove strategies have made him well known.In this work we want to concentrate on the evolutionary stability. Nash equilibrium doesn't imply stability that means if we perturb a little bit in the components of the vector a Nash equilibrium we may end up with a new Nash equilibrium. But in our case we need the clusters should be stable. In Evolutionary Game Theory,they introduce a refinement of Nash equilibrium that is Evolutionary

Stable Strategy which is stable and that always implies a Nash equilibrium.

Evolutionary stable strategies

Considering a large population whose chosen players are playing the same strategy, if small mutation choose playing a different strategy would die out, the strategy is called evolutionarily stable strategies. large population of players, there is a set of pure and mixed strategy set from which the selected two players choose their strategies which can be pure or mixed. The mixed strategy set is described as a standard simplex $\Delta = \{x \in R_+^n : x \geq 0, e^T x = 1\}$ where P is the pure strategy set ,of size n, for each players.The payoff associated to the two randomly chosen players is denoted by $u(x_1, x_2)$ & $u(x_2, x_1)$ when the players choose the strategies x_1 & x_2 from the set of the strategies in the standard simplex. The payoff matrix M is of size $n \times n$ and so the payoff $u(x_1, x_2) = x_1 M x_2$.Consider the players in the population that choose the same strategy ,$x \in \Delta$, called incumbent strategy and small number of players appear as a mutant and choose a strategy ,$y \in \Delta$ called mutant strategy. If $\epsilon \in (0, 1)$ is the sahre of the mutants which appeared in the population, a mutant and an incumbent have a probability of ϵ and $1 - \epsilon$ to be drawn randomly from the population with the appeared mutant.That means the population mix is ϵ & of y and $1 - \epsilon$ of x.

Example: Let's see if the cooperation and the defect,of the prisoner's dilemma, are Evolutionary Stable strategies.

	Cooperate	Defect
Cooperate	2,2	3,0
Defect	3,0	1,1

Is a cooperation or Defect evolutionary stable?

Let's assume that cooperation is evolutionary stable that means everyone in the population is adopted to play cooperation(referring x) and that there is a small mutation adopted to play defect(refer y). The population mix is then ϵ of y which is defect in this case and $1-\epsilon$ of x which is cooperate. During the random matching each of the players paired with $1-\epsilon$ chance to the cooperator and with ϵ chance to the non cooperator. In this case the cooperators are the incumbents and they have the average payoff of

$$(1-\epsilon)*2+\epsilon*0$$

And the mutants are with an average payoff of

$$(1-\epsilon)*3+\epsilon*1$$

which is better, so this mutant survive which implies that cooperation is not an evolutionary stable. If we do same for the defect we can see that it is an evolutionary stable strategy.

Evolutionary stable strategies imply Nash equilibrium,but not all strategies that are Nash are Evolutionary stable. We can also say ,for strategy a, that (a,a) is not a Nash equilibrium implies 'a' is not evolutionarily stable, but if 'a' is evolutionarily stable then it implies (a,a) is a Nash equilibrium. Consider the following two examples for the the Nash equilibrium and Evolutionary stability.

Example:

	a	b	c
a	4,4	0,0	0,0
b	0,0	0,0	1,1
c	0,0	1,1	0,0

Referring the above payoff matrix, suppose all players in the population adopted to play strategy c so that the population mix is $1-\epsilon$ of c (incumbent) and ϵ of b (mutant). The average payoff to incumbents adopted to play c is $(1-\epsilon)$ * 0 + ϵ*1 and that of the mutant is $(1-\epsilon)$ * 1 + ϵ*0 which shows 'b' is doing better than strategy 'c' which implies that a mutaant b's will not die out if it intrudes population with all players adopted to play a strategy 'c'. So a a population which consists of fully c's or b's can't be evolutionary stable as they can be intruded by a mutant b's or a mutant c's respectively. We can also see that 'c' is not dominated but playing against itself ,(c,c), is not a Nash equilibrium.

Example: Does a Nash equilibrium mean evolutionary stable?

	a	b
a	4,4	0,0
b	0,0	0,0

Does a Nash equilibrium mean evolutionary stable?

Here we can see that eventhough both (a,a) and (b,b) are Nash equilibrium, b is not evolutionary stable.

Formal definitions:

Definition: A strategy x $\in \Delta$ is evolutionary stable strategy if there exists a mutation of size $\bar{\epsilon}$ that are adopted to play strategy y $\neq$ x and satisfy the following inequality for any $\epsilon \in (0,\bar{\epsilon})$

$$(1-\epsilon) \textbf{ * u(x,x)} + \epsilon\textbf{*u(x,y)} > (1-\epsilon) \textbf{ * u(y,x)} + \epsilon\textbf{*u(y,y)}$$

Definition: A Nash equilibrium is an Evolutionary Stable Strategy(ESS) if for all strategies y

$$y^T Ax = x^T Ax \text{ implies } x^T Ay > y^T Ay.$$

We can express this definition as follows. Suppose x is a Nash equilibrium and suppose that y is another Nash equilibrium, if I get the same score when y is played against its opponent and when when x is played against itself, by changing the role of x and y we get $x^T Ay > y^T Ay$. If y is a strategy which is as good as x and if I change the role of the players and if my opponent is playing y, I better choose playing x against y rather than y. This is the condition that allows Nash Equilibrium to be stable and resist any small perturbation in the components of the vector. So the notion of Evolutionary Stable Strategy is the notion that we are looking for for our clustering purpose. Evolutionary Stable Strategies satisfies both the internal and external criteria of clustering. When we come to our optimization problem in doubly symmetric games i.e $A = A^T$ we have the notion of Nash Equilibrium coincides with the notion of Evolutionary Stable Strategies in the standard simplex Δ. While Nash Equilibrium is local maximizer of $x^T Ax$, the Evolutionary Stable Strategies is the strict local maximizer of $x^T Ax$. Here nothing can prevent us from giving the same definition of symmetric case of the dominant set for the non symmetric one, and so we can say the notion of Evolutionary Stable Strategy coincides with the notion of dominant set. In the next chapter we will see this coincidence the definitions and how the internal and external criterion of a cluster are satisfied.

2.3 Multiobjective Games

Many practical problems in this real world are not with a single objective,rather they are with multi-objectives, so it is better to think about multi-objectives than the con-

ventional game theoretic problems.So in many practical scenarios, single objective case become handicapped in recognizing the different relations between various measures (the similarities),that the players perceive in an evolutionary sense.In this case we need a vector payoff function which is the only difference with that of the single objective game i.e. if O = 1,2,...o is the set of the objectives, and all the players have the same payoff as M= $(M_j)_{j\in O} = (M_1, M_2, ..., M_o)$, where each M_i i ∈ O is an n × n matrix. When the players (mutant & incumbent) are drawn to play a game, their payoff is given respectively as:

$$\text{u(x,y)} = \mathbf{x}\text{M}\mathbf{y} = (\mathbf{x}M_1\mathbf{y}\ ,\ \mathbf{x}M_2\mathbf{y}, ...\mathbf{x}M_o\mathbf{y},) \in \mathbb{R}^o$$
$$\text{u(y,x)} = \mathbf{y}\text{M}\mathbf{x} = (\mathbf{y}M_1\mathbf{x}\ ,\ \mathbf{y}M_2\mathbf{x}, ...\mathbf{y}M_o\mathbf{x},) \in \mathbb{R}^o$$

Chapter 3

Dominant Sets

3.1 Introduction to dominant Sets

The notion of dominant set was introduced for the first time by Pelillo and Pavan [PP03b] in 2003. It is actually a combinatorial concept in graph theory that generalize the notion of the maximal clique to an edge-weighted graph. clustering is one of the effective ways for identifying and extracting different groups (compact set) which are similar to each other but different from the other groups (In clustering groups indicate highest similarities with in the cluster than the elements from the outside cluster) .Clustering problem appears in many areas like image processing, computer vision, bioinformatics, signal processing , medical imaging and others.The aim of clustering problem is to partition the given input, the set of n objects that can be arranged as an n by n matrix, in to different similar groups that satisfy some conditions. Depending on the format of the input objects we have two variations of clustering problems: Central (feature based) and Pairwise (graph based). For the first case the input objects are given to the algorithm are represented interms of feature vectors, so each object is represented as a point in an n dimensional space so that we can calculate some similarities between the points e.g using Euclidean distance. Some of the algorithms of this type are K-Means, Mean Shift and others. There are however applications where the feature based representation is not easily determined, situations like when the objects to be clustered are given as graph representation. In these cases we have to find the similarities between graphs, and there

are ways to compute similarities between graphs. Even though it is NP-hard to compute the similarities between arbitrary graphs, some instances of graphs can be computed in polynomial time. So even though we can't extract the feature based representation, it is possible to get the similarity of objects. Pairwise clustering algorithm is applied in this case. This algorithm accepts as an input a matrix of similarities and it tries to partition the data based on certain coherency criterion. This algorithm as I wrote above it just take as an input the similarity matrix, it doesn't know how we represent the data: It can be vectors, it can be graph or it can be other. So this algorithm is general than the other one as any representation is possible.

Let's try to answer this question **What is A Cluster?** The notion of a cluster is defined on the bases of two criteria[PP03b]: **Internal criterion** which deals with the similarity of objects of the inter cluster and the **External criterion** that deals with with the dissimilarity of the objects outside the cluster and the inside one. So informally , a cluster should satisfy the above two criteria. We can see in the following figure3.1 we have two clusters and the objects that are bounded using the curve can't form a single cluster as the group even though it satisfies the first criterion it doesn't satisfy the second criterion since it is contained in a larger coherent group.Here is the starting point for the definition of the dominant set.

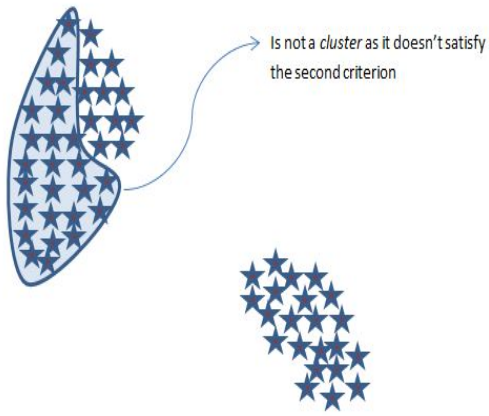


Fig. 3.1: Cluster example

3.2 Dominant Sets Clustering

Usually when we deal with the pairwise clustering processes, we represent the objects to be clustered as an undirected edge weighted graph where the n given points are represented by the vertices, and our similarities are the the weights of the neighbour similarities (edges). This graph is then represented as an n by n similarity matrix where the value of the matrix are the weights that determines the corresponding similarity of the points of the corresponding column and row. That is if our similarity matrix is W, then the value of $w_{i,j}$ represent the similarity between the vertex i and the vertex j(which is the edge weight). Since there is no an edge that connect a vertex to it self the main diagonal of the matrix is set to zero.

If we start from a very simple case, the binary case, our matrix become a 0,1 combination matrix that means an intermediate value is not allowed for the similarity(either they are similar or dissimilar). Here the graph is an undirected unweighed graph,so what is the sort of structure in this graph that satisfy both the internal and external criteria? The answer is a very classic notion of graph theory which is the notion of a `Maximal`

`Clique`. It is better to remember the definition of clique and see its meaning in a graph.

Mutually adjacent vertices form a set of `clique` and `clique` which is not part of a larger clique becomes a `Maximal Clique`.There are also the notions of a a `Maximum Clique` and a `Strictly Maximal Clique` that comes from the need of the stability of the set.

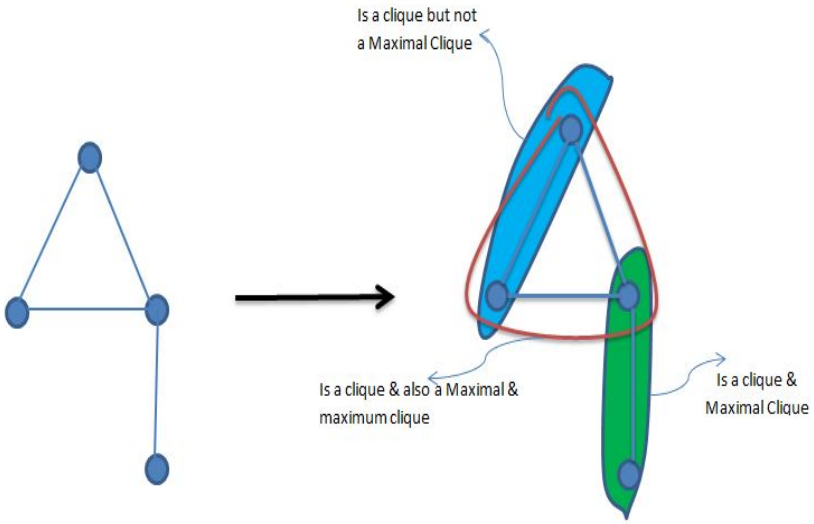


Fig. 3.2: Cluster example

A clique is called a `maximum clique` if it has the largest cardinality whereas For a `clique` to be `strictly maximal clique`(which is stable set) it should have the following property. If a graph is a maximal clique, all the vertices outside it can't have a number of edges,incident on its vertices, which is more than one less the cardinality of itself. Strictly maximal cliques are important as maximal cliques are sometimes unstable, which means if one vertex is dropped from the maximal clique and an external vertex is added, a new maximal clique will emerge. We can see this instability in the following figure3.3:

When we have a binary similarity, the notion of a cluster coincides with the notion of maximal clique, means it satisfies both criteria.However this doesn't work with the edge weighted graph. So, how can we generalize the notion of a maximal clique into an edge

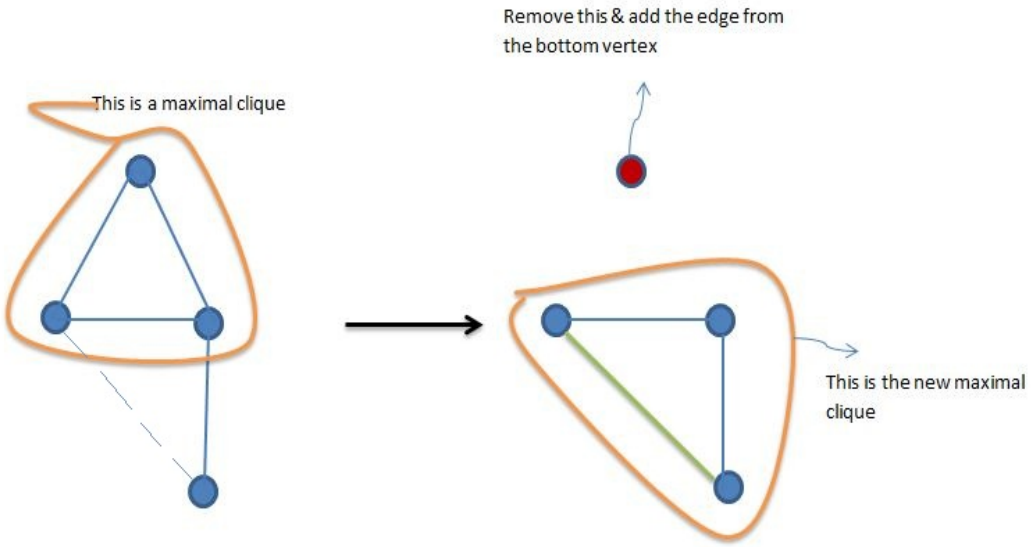


Fig. 3.3: Cluster example

weighted graph. Here comes the notion of dominant set into the picture.When you think about dominant set(a combinatorial concept in graph theory), the main objective is to generalize the notion of the maximal clique to an edge weighted graph.Before the main definition of dominant set, given by Pelillo and Pava [PP03b], let's see some definitions and some ideas that leads them to the main definition of the notion of the dominant set.

Definition: The sum of the weights(the similarities) of the edges that connects the point to all other points in the set divided by the cardinality of the the set(this makes the average) gives us the average weighted degree of a point(AWD). Mathematically the average weighted degree of a point ip with respect to a set of vertices S that is not an empty set is expressed as

$$AWD_S(ip) = \frac{1}{|S|} \sum_{p \in S} w_{ip,p}$$

Where $w_{ip,p}$ is the weight(the similarity) between the two points

This is the average weighted similarity between the point p and the other points in the set S.

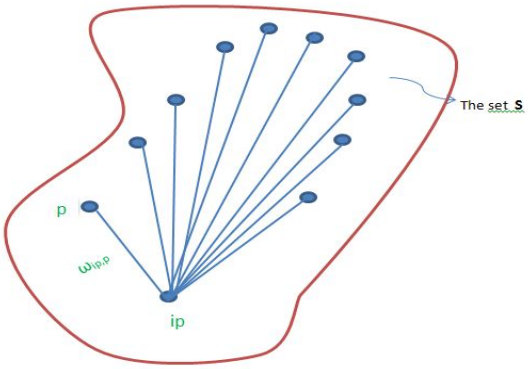


Fig. 3.4: Average weighted degree of point ip

The relative similarity,$\phi_S(ip, op)$, between two objects,ip and op(i and o to indicate the points inside and outside), with respect to the average similarity between node ip and its neighbours is described as the difference between the absolute similarity between ip and op ($w_{ip,op}$) and the average weighted similarity $AWD_S(ip)$

$$\phi_S(ip, op) = w_{ip,op} - AWD_S(ip)$$

This $\phi_S(ip, op)$ can be positive or negative based on the value of the absolute similarity and the average weighted similarity. If the absolute similarity is greater than the average weighted similarity, it is positive, otherwise it becomes negative.

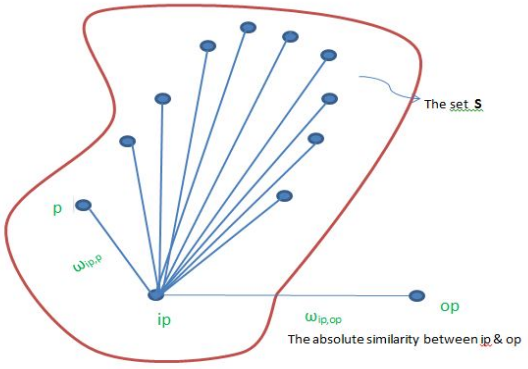


Fig. 3.5: Relative similarity between two objects

Using this result of $\phi_S(ip, op)$ it is possible to have the following recursive definition that allows us to assign a weight to a node. This is the main definition that allows to give

the main definition of dominant set. If the cardinality of the set S is 1 then by definition $W_S(ip) = 1$. Otherwise we have to sum up all the relative similarities between i and all other points in the set S, and this tells us how similar is point ip with respect to all other points in the set S.

$$W_S(ip) = \sum_{p \in S \setminus \{ip\}} \phi_{S \setminus \{ip\}}(ip, p) W_{S \setminus \{ip\}}(p)$$

Then the weight of the set S is the sum of of each weights $W_S(ip)$. We know $W_S(ip)$ the measure of how much tightly a vertex is coupled with other set of vertices.Let's see some numerical examples before the main definition of dominant set.

Example: Calculate the weight of following graph with 3 vertices.

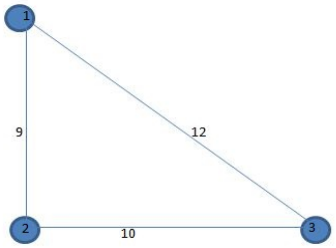


Fig. 3.6:

Solution: Before doing anything we can know the vertex with a highest weight and the lowest weight. We can do this simply by adding the edge-weights that points to the vertices. So 10+12=22 > 9+12=21 > 9+19=19. This implies that $W_{\{1,2,3\}}(3) > W_{\{1,2,3\}}(1) > W_{\{1,2,3\}}(2)$ Let's do the calculation using the formula and compare the result with this one.

we have to remember these points

$$W_{\{i,j\}}(i) = W_{\{i,j\}}(j) = \omega_{i,j}$$

$\phi_{\{i,j\}}(j) = \omega_{i,j}$

$W_{\{i\}}(i) = 1$ by definition when the cardinality is one

$$\begin{aligned} W_{\{1,2,3\}}(1) &= \phi_{\{2,3\}}(2,1)W_{\{2,3\}}(2) + \phi_{\{2,3\}}(3,1)W_{\{2,3\}}(3) \\ &= (\omega_{2,1} - AWD_{\{2,3\}}(2))\omega_{2,3} + (\omega_{3,1} - AWD_{\{2,3\}}(3))\omega_{2,3} \\ &= (9 - 10/2)10 + (12 - 10/2)10 \\ &= 110 \end{aligned}$$

Similarly we get $W_{\{1,2,3\}}(2) = 84$ and $W_{\{1,2,3\}}(3) = 117$. So the total weight is the sum off all these three which is $110 + 84 + 117 = 311$.Now let's see the effect of adding an external vertex to this graph. What will be the sign of the weight of the new vertex, meaning what is the overall similarity of this new vertex with respect to the previous vertices.What is the sign of the weight of these new added vertices(vertex 4 and 5) in the figure below?

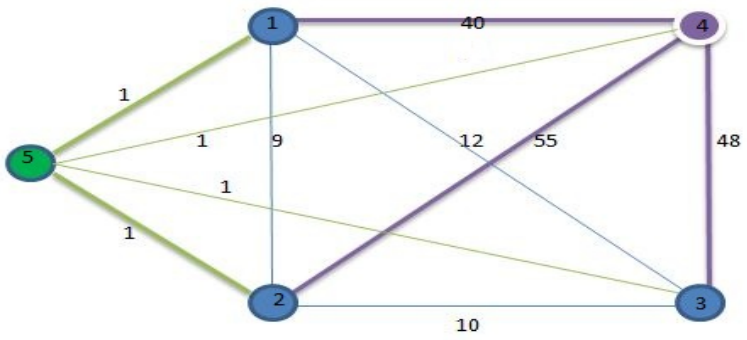


Fig. 3.7:

It is clear to see that $W_{\{1,2,3,4\}}(4) > 0$ as the heaviest points are pointing to this vertex

so it is highly tightly bonded or grouped to other vertices, where as $W_{\{1,2,3,4,5\}}(4) < 0$ as it is the opposite of the first case.Now we can say we are in a time to see the main definition of dominant set.

Definition: Pelillo and Pavan [PP03b]: A nonempty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any nonempty $T \notin S$, is said to be dominant if:

1. $W_S(i) > 0$ for all $i \in S$
2. $W_{S\cup\{i\}}(i) < 0$ for all $i \notin S$

These two conditions are exactly the same condition of the clustering criteria, so both criteria for a cluster are satisfied. Here we can say the notion of a cluster coincides with the notion of dominant set. If we know they coincide, how can we calculate the dominant set? or how can we partition a set of data in to dominant set? Pelillo and Pavan, instead of using a standard algorithm to find dominant set,they transform the purely combinatorial problem of finding a dominant set in a graph in to a pure quadratic optimization problem and use evolutionary game theory dynamical system to solve the optimization problem.Using this algorithm it is possible to select out the identified dominant set from the graph and continue until the stopping criterion which checks if we have an empty set of vertices.

3.3 Dominant Set Clustering Example

Let's see a simple example with 10 points in a two dimensional space of the X-Y plane that forms two clusters. Here are our points as a vector of points and there representation in 2D space.

```
1.0000    1.0000    2.0000    2.0000    3.0000    3.0000    4.0000    4.0000    1.5000    3.5000
1.0000    2.0000    1.0000    2.0000    5.0000    6.0000    5.0000    6.0000    1.5000    5.5000
```

Fig. 3.8: Points as a vector

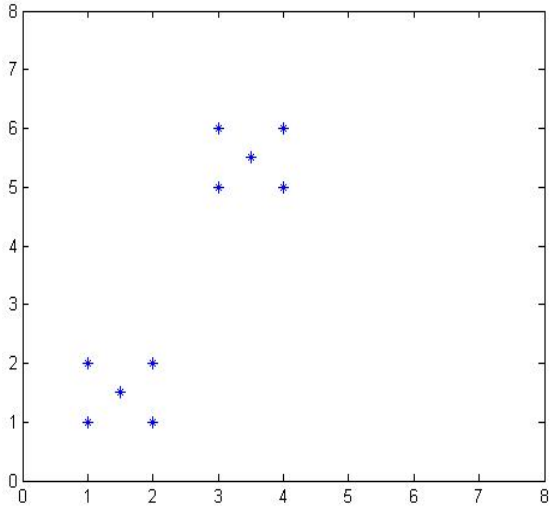


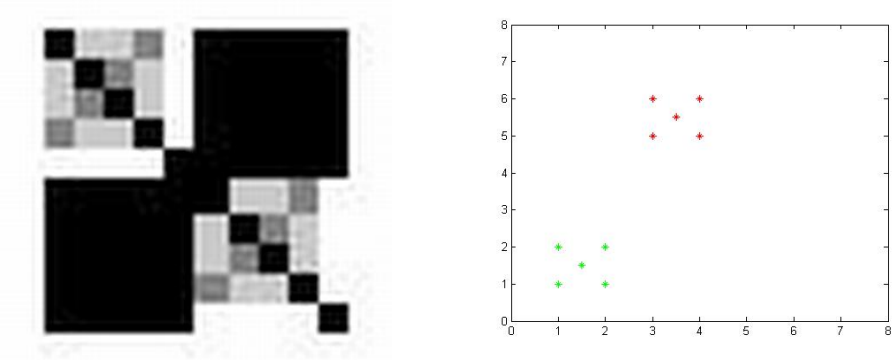
Fig. 3.9: Vector representation

This vectors are then changed to a distance matrix which is a matrix of the euclidean distance between each points. For the distance matrix D, $d_{i,j}$ represents the distance between point i and point j of the vector.Here is the distance matrix for our example.

0	1.0000	1.0000	1.4142	4.4721	5.3852	5.0000	5.8310	0.7071	5.1478
1.0000	0	1.4142	1.0000	3.6056	4.4721	4.2426	5.0000	0.7071	4.3012
1.0000	1.4142	0	1.0000	4.1231	5.0990	4.4721	5.3852	0.7071	4.7434
1.4142	1.0000	1.0000	0	3.1623	4.1231	3.6056	4.4721	0.7071	3.8079
4.4721	3.6056	4.1231	3.1623	0	1.0000	1.0000	1.4142	3.8079	0.7071
5.3852	4.4721	5.0990	4.1231	1.0000	0	1.4142	1.0000	4.7434	0.7071
5.0000	4.2426	4.4721	3.6056	1.0000	1.4142	0	1.0000	4.3012	0.7071
5.8310	5.0000	5.3852	4.4721	1.4142	1.0000	1.0000	0	5.1478	0.7071
0.7071	0.7071	0.7071	0.7071	3.8079	4.7434	4.3012	5.1478	0	4.4721
5.1478	4.3012	4.7434	3.8079	0.7071	0.7071	0.7071	0.7071	4.4721	0

Fig. 3.10: Distance Matrix representation

Now from this distance matrix it is possible to create some similarity matrix. For example using this formulation, $exp^{-(d^2/\sigma^2)}$,we can get the weight,which is inversely proportional to the corresponding distance measure, that represent the similarity between two points. 'd' is the distance measure between two points, which is the elements of the distance matrix and σ is the parameter that determines how much the weight is decreasing. This similarity matrix is given as an input to the algorithm that calculates the dominant set.



(a) (b)

Fig. 3.11: (a)The similarity matrix passed to the algorithm. (b)The resulting cluster

3.4 Dominant Set and the Quadratic Optimization

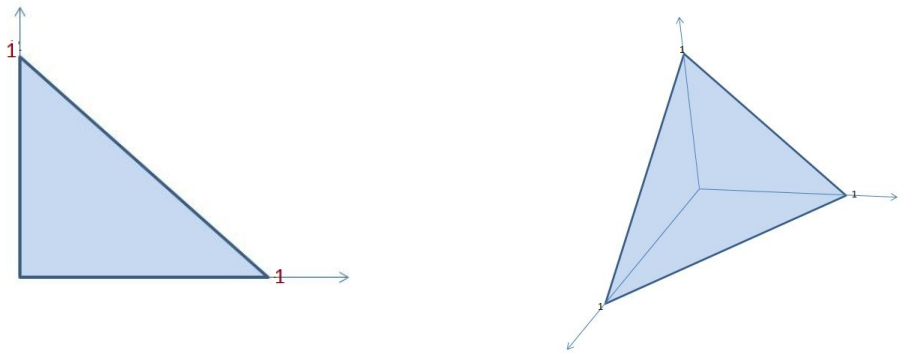
Here, since we have now the notion of dominant set, we are going to start the math to calculate it so that we can partition the given set in to clusters.Even though it is NPhard to find the maximum clique in a graph we can calculate the maximal clique simply in a quadratic time.As I wrote before Pelillo and Pavan, instead of using a standard algorithm to find dominant set,they transform the purely combinatorial problem of finding a dominant set in a graph in to a pure quadratic optimization problem and use evolutionary game theory dynamical system to solve the optimization problem.This problem is the general form of the well known problem from graph theory,Motzkin-Straus problem[MS65]. Lets see what this theorem says before the problem that generalize it.Actually it is a problem that creates the one to one correspondence between the maximal clique in the undirected un-weighted graph(binary case) and the quadratic optimization problem.

Motzkin-Straus theorem says given an undirected un-weighted graph G=(V,E) and W is the adjucency matrix of the graph.There is a one-to-one correspondence between the clique number of the graph $\omega(G)$(the value of the cardinality of the largest clique) and the maximal optimizer of the problem:

$$\begin{aligned} &maximize && \mathbf{f(x) = x^T W x} \\ &\text{subject to } x \in \Delta(thestandardsimplex). \end{aligned}$$

If $\mathbf{x}^*$ is the maximizer, then $\omega(G) = \frac{1}{1-f(x^*)}$

The standard simplex Δ is a simple geometrical structure. It is the intersection of the plane of equation sum of all x_i is one over the plane of the positive orthant,means all x_i should be grater than or equal to zero. Here is the graphical representation of the standard simplex when n=2 and n=3.



(a) (b)

Fig. 3.12: (a)The standard simplex when n=2 (b)The standard simplex when n=3

Other peoples,that try to solve clustering problem using different approach,e.g S. Sarkar and K. L. Boyer[SB98] use similar but fundamentally different quadratic program. The main idea they used,in spectral clustering problem,is to find the largest eigenvalue and the associated eigenvector of the similarity matrix W. They found this by maximizing this quadratic program

$$\begin{aligned} &\text{maximize} && \mathbf{x^T W x} \\ &\text{subject to } x \in S(theSphere). \end{aligned}$$

This means $x^T x = 1$

As we see the objective function is the same but the domain is completely different.This

method has some drawbacks like it considers only the positive eigenvalues and also finding the maximum value is NPhard since it is simulated to the maximum clique.

Now let's see one definition before the theorem which establishes a one-to-one correspondence between dominant sets and the strict local maximizer of the quadratic program $x^T W x$ over the standard simplex.This theorem is the the theorem that generalizes the Motzkin-Straus problem.

Definition: A weighted characteristic vector is a vector in the standard simplex that can be defined as follows:

$$x^S = \begin{cases} \frac{w_S(i)}{W(S)} & if\ i \in S; \\ 0 & otherwise. \end{cases}$$

Theorem,Pelillo and Pavan: If S is a dominant subset of vertices, then its weighted characteristic vector x^S is a strict local solution of objective function in the standard simplex.

Conversely if x^* is a strict local maximizer of the objective function in the standard simplex then its support

$$\sigma = \sigma(x^*) = \{i \in V : x_i^* \neq 0\}$$

is a dominant set provided that $w_{\sigma \cup \{i\}} \neq 0 \quad \forall i \neq \sigma$

The weight characteristic vector is a vector of as many elements as there are vertices in the graph where the i^th component will be zero if $i \notin S$, otherwise it takes the ration as shown above in the theorem. If we sum up all its elements,we will end up at a point in the standard simplex.As we can see all the components of the vector are positive as they are the ratio of two positive numbers that are weights and the sum of all the

components give for us one.So if S is the dominant set, this vector is the strict local maximizer of $x^T Wx$ in the standard simplex. On the other way if we are given that x^* is the strict local maximizer of the objective function in the standard simplex,take the support of the vector which is the indices corresponding to positive components of the vector and these represents subset of vertices in the graph that corresponds to the dominant set. Now we are in a point that dominant sets are characterized as a continuous form optimization problem. So after this we have to think about the way that we have to solve the optimization problem and get the dominant sets.There are many ways to solve the optimization problem. Here we are going to use the dynamical system equation,the replicator equation, that comes from the evolutionary game theory2

In Evolutionary Game Theory,they introduce a refinement of Nash equilibrium that is Evolutionary Stable Strategy which is stable and that always implies a Nash equilibrium.

Definition: A Nash equilibrium is an Evolutionary Stable Strategy(ESS) if for all strategies y

$$y^T Ax = x^T Ax \text{ implies } x^T Ay > y^T Ay.$$

We can express this as follows. Suppose x is a Nash equilibrium and suppose that y is another Nash equilibrium, if I get the same score when y is played against its opponent and when when x is played against itself, by changing the role of x and y we get $x^T Ay > y^T Ay$. If y is a strategy which is as good as x and if I change the role of the players and if my opponent is playing y, I better choose playing x against y rather than y. This is the condition that allows Nash Equilibrium to be stable and resist any small perturbation in the components of the vector. So the notion of Evolutionary Stable Strategy is the notion that we are looking for for our clustering purpose. Evolutionary Stable Strategies satisfies both the internal and external criteria of clustering. When we come to our optimization problem in doubly symmetric games i.e $A = A^T$ we have the notion of Nash Equilibrium coincides with the notion of Evolutionary Stable Strategies

in the standard simplex Δ. While Nash Equilibrium is local maximizer of $x^T Ax$, the Evolutionary Stable Strategies is the strict local maximizer of $x^T Ax$. Here nothing can prevent us from giving the same definition of symmetric case of the dominant set for the non symmetric one, and so we can say the notion of Evolutionary Stable Strategy coincides with the notion of dominant set. This means that even-though we have now a non symmetric or even a negative matrix A, we can give the same definition of the average weighted degree(AWD), the relative similarity Φ, the weight and the total weight that already gave for the symmetric matrix.

Theorm Pelillo and Pavan: If I take the an Evolutionary strategy, then I take the support of the set and this is a directed dominant set. Or we can say if I take the dominant set and if I set a characteristic vector this is an Evolutionary Stable Strategy

So if I know there is a one to one correspondence between the ESS and dominant set, how can I calculate this directed dominant set, that means how can I calculate this ESS? I mean we are now at a point at which we are no longer rely on the fundamental theorem of natural selection, we are not doing or solving an optimization or maximization problem, instead we are looking for an equilibrium of a game so there is no in general any objective function to optimize here. The replicator dynamics work on a non symmetric and on a negative matrix. So here it is not possible to say the algorithm is maximizing or minimizing something at each step, instead one can say the following two points: From the theorem below we can see there is a one to one correspondence between the Nash equilibrium and the limit point of the trajectory which starts from the interior. This means if one start from the replicator dynamics from the interior point, when it converges it is the Nash equilibrium. The other point that we can say is about the ESS i.e. if a point in the standard simplex is ESS then it is asymptotically stable. This is to mean that around the point there is a basin of attraction so if the replicator dynamics starts with in this basin of attraction it will converge towards it. So in general we can

say when the algorithm converges, usually it is an ESS and we can say it is necessarily a Nash equilibrium. Here is the generalized theorem.

Theorem:Pelillo & Pavan: A point $x \in \delta$ is the limit of a trajectory of the replicator dynamics starting from the interior of the standard simplex iff x is a Nash equilibrium. Further, if point $x \in \delta$ is an ESS then it is asymptotically stable.

Chapter 4

Our contribution

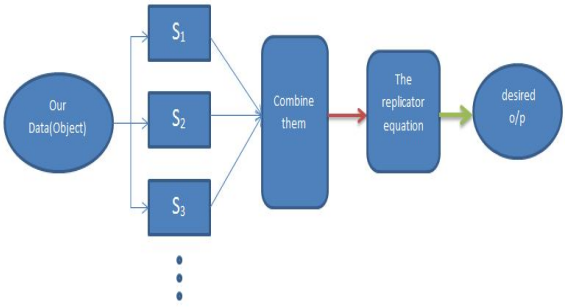
4.1 Introduction

Starting from the past many researchers have been trying to solve problems like multiple similarities.To solve this type of problems peoples usually transfer the data to a different form, they represent the data in different form. It is very important to represent the data in a way we can solve the problem easily.When we map our data to an appropriate feature space, it is very important to choose the right similarity measure that represents our data as it has a great impact on the performance of an algorithm that solves the problem.Thinking that combining different similarity measures in to one to represent the data help solve problems effectively, people create different representations for a data and combine in different way to use it as a representation of the data. Since most of the time it is very difficult to know the best representation of our data it is better to do combining different measures.Whether it is combined or not almost all the works that are done to solve such problems as specific in the sense that they use a very specific similarity measures. For examples in the papers [KZ03] , [HCHC11], [?] used only a positive semi-definite matrices. Here in our case we are working in a generalized version.To make this representation general we use the concept of dominant set and evolutionary game theory. As it is written in the paper refereed in [PP03b] and [PP03a] Dominant sets are a new graph-theoretic concept that has proven to be relevant in clustering as well as image segmentation problems.In those papers Pelillo

and Pavan referring the problems in dominant set and giving the solution,a new hierarchical approach to clustering is suggested , which is based on the idea of properly varying the regularization parameter during the clustering process.At each level of the hierarchy,they found the solutions of the quadratic programs using a dynamical system from evolutionary game theory.It is known that algorithms in Graph-theories are used to search for certain combinatorial structures in the similarity graph. What they gave attention was on the complete-link algorithm [JD98] which search for a complete subgraph,namely a *clique*. Addressing that the notion of maximal clique which defined on unweighted graphs is infeasible,even-though it is possible to generate a hierarchy of clusters using a threshold operation [JD98] but the size of the data to be clustered is too large, they extend the notion of the maximal clique to edge-weighted graph, using the notion of dominant sets, which allow them for the development of a new nonhierarchical (partitional) clustering approach.The strong relation between dominant sets and the extrema of a quadratic form over the standard simplex allows them to find dominant sets using a dynamical systems arising in evolutionary game theory [J.W95].We are using this methodology for multiple similarity using multi-objective sense. Many practical problems in the real world are not with a single objective, so it is better to think about multi-objectives than the conventional game theoretic problems.So in many practical scenarios, single objective case become handicapped in recognizing the different relations between various measures,the similarities,that players perceive in an evolutionary sense.In this case we need a vector payoff function[3].Similarly in this work we are going to combine the different similarity matrices and find the dominant set in the combined similarity measure.

4.2 The Proposed Approach

Let n be the number of objects in the data and lets assume we can have m different similarity matrices for the data, the proposed approach combines all these similarities in to a single similarity measure and apply on it the replicator dynamics to find the dominant set. The proposed approach to learn pairwise similarity based on multicriteria clustering is schematically presented in figure 1. S_i represents the different similarity matrices, and the combined version of these matrices is given as an input to the dynamical systems to get the desired output,the dominant sets.



4.2.1 How we combine

In most of the studies, each partition is given an equal weight in the combination process and all similarities in each partition contribute to the combined solution equally. Selection criteria amongst possible combination solutions has been proposed based on the importance of the similarity measure.One of the most important techniques in vector optimization is scalarization.If we have q objectives,the *weighted sum scalarization*,the method that we used to combine, will have the following form: $\sum_{i=1}^{q} wA_ix = w^TA_ix$ where $w \in \Delta$,the standard simplex $\Delta = w \in R^n : w >= 0, e^Tw = 1$

Somasundaram and John S. Baras [M.] show the equivalence of the player's best strat-

egy against opponents to a parametric multi-objective linear program and also they show that this relation is useful in constructing the tradeoff weights among the various payoffs,objective similarities. For $\mathbb{R}^l_{>0}$ denote the set of l-dimensional vectors such that for $a \in \mathbb{R}^l_{>0}$ $a > 0$, the weighted sum linear program, for $w \in \mathbb{R}^l$ can be written as:

$$\max_{x \in \Delta} w^T U x. \qquad (1)$$

Where each row of U in our case represents the different similarity measures. We can define the row of U as follows:

$$r_1 = x^T S_1$$
$$r_2 = x^T S_2$$
$$r_3 = x^T S_3$$
$$.....$$

After we get the desired tradeoff weights,solving the above optimization problem(problem 1), we combine our different similarity measures and give as an input to the replicator dynamics so that we can get the selected dominant set.If we say our combined similarity measure is M, it can be written like this

$$M = \sum_{i=1}^{k} w_i S_i$$

Where k is the number of different similarity measures that we have and w_i is the weighted associated to the i^{th} similarity measure.

4.3 Dominant Sets and Replicator Dynamics

In Graph-theoretic definition of a cluster [PP03b] the definition of both a cluster,weighted characteristic vector and dominant sets are well defined. In this work we are also using

the same approach. If we have an edge-weighted graph $G = (V, E, w)$ and its weighted adjacency matrix A, consider the following quadratic program (which is a generalization of the so called Motzkin-Straus program [MS65]):

$$\begin{aligned} &\text{maximize } f(x) = \tfrac{1}{2} x^T A x \\ &\text{subject to } x \in \Delta \end{aligned} \qquad (2)$$

a point x^* is said to be a local solution of problem (2) if there exists an $\epsilon > 0$ such that $f(x^*) \geq f(x)$ for all $x \in \Delta$ whose distance from x^* is less than ϵ, and if $f(x^*) = f(x)$ implies $x^* = x$, then x^* is said to be a strict local solution.And there is a theorem,with its proof, written in [PP03b] which says

Theorem 1:If S is a dominant subset of vertices, then its weighted characteristics vector x^s is a strict local solution of program (2).

So from the theorem it is possible to say dominant sets are in correspondence with (strict) solutions of quadratic program. This theorem can made them conclude its provision of a tight correspondence between the problem of finding dominant sets in an edge-weighted graph and that of finding solutions of a quadratic program. By virtue of this theoretical result, they were able to find a dominant set by first localizing a solution of program (2) with an appropriate continuous optimization technique, and then picking up the support set of the solution found. To solve problem (2), it is possible to use both a continuous or discrete for of the dynamical system from evolutionary game theory. Let M be a non-negative real-valued $n \times n$ matrix, here we can have both type of the dynamical system as follows:

$$\dot{x}_i(t) = x_i(t)[(M\mathbf{x}(t))_i - x(t)^T M\mathbf{x}(t)] \qquad (3)$$

$$x_i(t+1) = x_i(t)\frac{(M\mathbf{x}(t))_i}{\mathbf{x}(t)^T M\mathbf{x}(t)} \qquad (4)$$

The dot in the continuous version represents the derivative with respect to the time, and for the discrete version of this continuous dynamical system we have the values at the time difference.This dynamics has an important characteristics in the simplex Δ.Any trajectory starting in Δ will remain in Δ for all the time in the future,and also the stationary points coincide and both can solve the following equation

$$x_i\ [(\mathrm{M}\mathbf{x}(\mathrm{t}))\ _i\ -\mathrm{x}(\mathrm{t})^T\mathrm{M}\mathbf{x}(\mathrm{t})] = 0$$

for i= 1,2,....n. This dynamical properties under the simplex help everyone solve problems of such type.

4.4 Experimets and Experimental Results:

We applied the prposed approach to a different real data sets from UCI: Iris, Seeds, Pima and Ionosphere and also to a toy data set generated by ourselves. When we apply the approach to these data sets we remove the lables to the data sets first and try to recognize the pattern in the data set and then compare with the ground truth to know the error done. The parameter σ varies depending to the characteristics of the data set. Here the following graph summarizes and show the error that we get when we apply our method to the toy data set(random samples with two clusters of the Gaussian distributions with some noise added to it) with 120 data points with 3 features.First the similarity matrices are calculated,using the euclidean distance and Gaussian, and then we applied the replicator dynamics on each similarity measures and also on the combined similarity matrix. The x- axis represents the experiments iteration points (the

first, second .... until the 10^{th} experiment) and the y-axis represents the error done during the operation.This experiment is done for 20 times and this is the average plot that we got. So in total the experiment is done for 200 times.

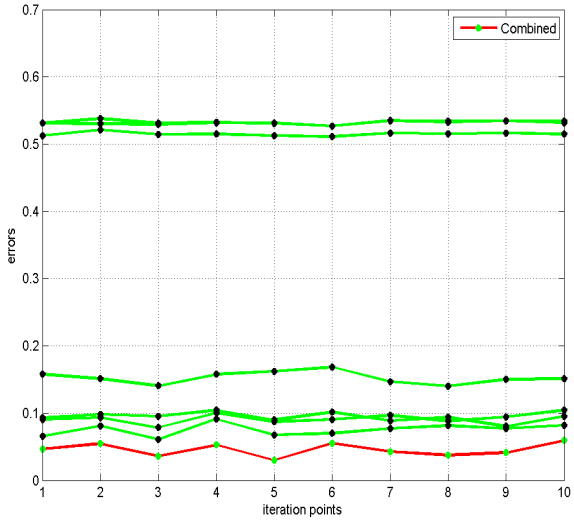


Table 1 summarizes the results that we get for the data sets: Iris,Seeds, Pima and Ionosphere.The experiment is done similarly as above and the table provides us a direct comparison between the different similarity measures and the combined version.We have done the experiment using different type of similarity matrices like Gaussian = $exp\frac{-D^2}{2\sigma^2}$, Linear = $P * P'$ and also Polynomial = $constant + P * P'$. The experiment is done for each similarity matrices and for the combined one, and the table shows us the resulted error. The result in the table is the average result of the experiments done repeatedly for 20 times. As can be seen from the table we can see the smallest error is from the last row which is the case of the combined similarity measure.

Bibliography

[EARJTAEL06] Boulogne T. El-Azouzi R. Jimenez T. Altman E. and Wynter L. *survey on networking games in telecommunications.Computers and Operations Research*. 2006. 16

[HCHC11] Yung-Yu Chuang Hsin-Chien Huang and Chu-Song Chen. Multiple kernel fuzzy clustering. 2011. 8, 40

[JD98] A. K. Jain and R. C. Dubes. Algorithms for clustering data. 1998. 41

[JS] Don H. Johnson and Sinan SinanoviÂ´c. *Symmetrizing the Kullback-Leibler distance*. 16

[J.W95] J.W.Weibull. Evolutionary game theory. mit press, cambridge ma. 1995. 41

[KP97] Daphne Koller and Avi Pfeffer. *Representation and solutions for game theoretic problems. Artificial Intelligence*. 1997. 16

[KZ03] Bin Zhao James T. Kwoky and Changshui Zhang. Multiple kernel clustering. 2003. 8, 40

[M.] Ehrgott M. Achieving symmetric pareto nash equilibria using biased replicator dynamics. iv, v, 42

[MM] William Pentney Marina Meila. *Clustering by weighted cuts in directed graphs*. 16, 17

[MS65] T.S. Motzkin and E.G. Straus. *Maxima for Graphs and a New Proof of a Theorem of TuraÂ´n*. 1965. 34, 44

[NR99] Nisan and Ronen. *Algorithmic Mechanism Design*. 1999. 17

[PP03a] M. Pavan and M. Pelillo. Dominant sets and hierarchical clustering. 2003. 40

[PP03b] M. Pavan and M. Pelillo. *A new graph-theoretic approach to clustering and segmentation*. 2003. 23, 24, 27, 31, 40, 43, 44

[SB98] S. Sarkar and K. L. Boyer. *Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors*. 1998. 35

[UMK]