A game theory framework for clustering

Abdellah Salhi[1], Berthold Lausen, Fajriyah Rohmatul, Marwa Baeshen, and Özgün Töreyen[2]

[1] University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK,
as@essex.ac.uk,
WWW home page: http://www.essex.ac.uk/maths/staff/profile.aspx?ID=1273
[2] ASELSAN, Defense Systems Technologies, Yenimahalle, 06172, Ankara, Turkey

Abstract. The Game Theory-based Multi-Agent System (GTMAS) of Töreyen and Salhi, [10] and [12], implements a loosely coupled hybrid algorithm that may involve any number of algorithms suitable, *a priori*, for the solution of a given optimisation problem. The system allows the available algorithms to co-operate toward the solution of the problem in hand as well as compete for the computing facilities they require to run. This co-operative/competitive aspect is captured through the implementation of the Prisoners' Dilemma paradigm of game theory. Here, we apply GTMAS to the problem of clustering European Union (EU) economies, including Turkey, to find out whether the latter, based on a number of criteria, can fit in the EU and find out which countries, if any, it has strong similaries with. This clustering problem is first converted into an optimisation problem, namely the Travelling Salesman Problem (TSP) before being solved with GTMAS involving two players (agents) each implementing a standard combinatorial optimisation algorithm. Computational results are included.

KEYWORDS: Multi-Agent System, Game Theory, Payoff Table, Optimisation, TSP, Clustering

1 Introduction

The clustering problem occurs in many areas and, when constrained, for instance by imposing that two objects be in the same cluster or otherwise, it is intractable, [20]. Following the work of Lenstra, [7], in which clustering was first mapped onto a TSP problem, we suspect as do others, [19], that the general clustering problem is also intractable. Clustering is the grouping or the partitioning of entities into subsets. There are many clustering techniques with different intuitions that can be put broadly in categories such as partitioning, hierarchical, density-based, grid-based, and model-based. To these, one can add more recent approaches such as neural networks, mixture-resolving and mode-seeking algorithms, nearest neighbour methods, fuzzy methods, evolutionary and search-based approaches, [6].

Almost always clustering algorithms require the number of clusters as a pre-specified input. However, it is usually not possible to know it *a priori*. To find out the best clustering with the optimum number of clusters, v-fold cross-validation [13] is used. It is a simple analysis that graphs the loglikelihood of the

total point-to-centroid distance to observe the cut-off number where increasing this number does not improve the distance as it did in the previous increases.

The case of interest here is that of European Union (EU) countries with the addition of Turkey. The issue is to find whether Turkey, both as an economy and a culture, can fit into the EU. Finding objectively how strong its similarities with EU countries are and with which ones it has similarities will potentially, we hope, reassure those concerned with this issue.

We intend to approach this clustering problem as an optimisation one by converting it first into a TSP. The solution approach is that described in [10, 12] and implemented as GTMAS. It uses two solution algorithms (solver agents) in the process, namely the Genetic Algorithm (GA), [5], and Simulated Annealing (SA), [11]. The solver agents play a game of the Iterated Prisoners' Dilemma (IPD) type, [1]. Appropriate payoff tables are used to encourage or otherwise competition for instance.

1.1 Background to the problem

The official negotiations process for Turkey to join the EU was launched in October 2005. To be a member state, Turkey should satisfy the Copenhagen criteria, which are five, [15]. They span mainly the political and economic sectors. But, cultural aspects are also considered. They are:

- **Political:** stability of institutions guaranteeing democracy, the rule of law, human rights and respect for and protection of minorities;
- **Economic:** existence of a functioning market economy and the capacity to cope with competitive pressure and market forces within the Union;
- **Acceptance of the Community Acquis:** ability to take on the obligations of membership, including adherence to the aims of political, economic and monetary union.

The French word "Acquis" may be translated into "Achievements".

Turkey's interest to join the EU goes back to 1959. Since then, policies within Turkey are adapted and implemented in order to meet the criteria. However, the accession process keeps running into difficulties.

There is a large body of literature reporting conflicting arguments, mainly in Politics and Economics outlets, on why the accession process has failed so far, [16, 18, 17]. Some believe that negotiations will not be concluded before 2027 with no guarantee of success. We have no intention to join the debate. Instead, we are after a more objective analysis based on societal data of EUROSTAT [4], [3] and TURKSTAT, [14]. These data represent a measurable aspect of the Copenhaguen criteria.

1.2 Factor selection

The following factors are selected to quantify societal similarities. They fall into four main categories: Population and Health, Living Standards, Education and Culture, and Work.

- **Population and Health**
  1. Population: Population of countries in 2006, [3].
  2. Marriages: Marriages in 2003 per 1000 persons [4], [14].
  3. Divorces: Divorces in 2002 per 1000 persons [4], [14].
  4. Male life expectancy at birth: 2002 data in years, [4, 14].
  5. Female life expectancy at birth: 2002 data in years [4, 14].
- **Living Standards**
  1. Gross Domestic Product (GDP): per inhabitant in 2006, [4].
  2. Serious accidents at work: from 1998 to 2002, [4].
  3. Passenger cars: Number per 1000 inhabitans in 2002, [4].
  4. Municipal waste: in kgs per person per year in 2003, [4].
  5. Length of railway lines: per 1000 km^2 in 2002, [4].
- **Education and Culture**
  1. Pupils and students: per 1000 inhabitants, [4].
  2. Number of foreign languages learnt by pupils in general education: in 2004, [4].
  3. Male individuals regularly using internet: Percentage in 2004, [4].
  4. Female individuals regularly using internet: % in 2004, [4].
  5. Cinema admissions: per inhabitant in 2006, [3].
  6. Religion: Religion is considered since it affects culture.
- **Work**
  1. Male unemployment rate: in 2004, [4].
  2. Female unemployment rate: in 2004, [4].
  3. Employed population in agriculture: Regular farm labour force per 1000 inhabitants in 2000, [4, 14].
  4. Researchers: Number per 1000 inhabitants in 2002, [4].

Note that a lot of the figures used have to be calculated from raw data found in [4, 14]. Raw data are normalised according to the standard normalisation,

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, \tag{1}$$

where z_{ij} denotes the normalised value of the i^{th} country with respect to the j^{th} feature, x_{ij} denotes the raw value of the i^{th} country with respect to the j^{th} feature, $\overline{x_j}$ denotes the mean for the j^{th} feature and σ_j denotes the variance of the j^{th} feature.

2 Modelling the clustering problem as a TSP

Clustering via solving a TSP has been studied by many, [7], [2]. The mapping from a rearrangement clustering instance to a TSP instance is particularly well defined in [8]; the objects to cluster are mapped to the cities in TSP and the dissimilarities between these objects are mapped to the distances between the cities. The objective of TSP, minimising the tour length, guarantees visiting the group of cities that are close to each other consecutively and tries to avoid big jumps to farther cities as much as possible. This is equivalent to the objective of clustering which is to group similar objects in the same clusters and non-similar ones in different clusters. In [8] it is concluded that TSP is equivalent to the problem of finding an optimal permutation, except that rearrangement clustering finds a path whereas TSP finds a cycle.

In [2] a method is given to find a path, or to find a cut-off point for the TSP tour. They introduce a dummy city which has the same distance C, to all of the original cities, where C is as small as possible. This mapping has been proved to be optimal, [2].

Lemma 1 [2]: *The direct distance between the two cities that are separated by the dummy city is greater than or equal to any of the distances between adjacent pairs of cities on the TSP tour, and the total distance of the TSP path is the smallest possible.*

This idea, [2], is then extended for the k-clustering problem, where k denotes the number of clusters. They introduce k dummy cities this time, which have a small distance C to every original city and an infinite distance to every other dummy city, in order to separate the TSP tour into k clusters. The optimality of this mapping is established by the following theorem.

Theorem 2 [2]: *With k dummy cities, the sum of the lengths of the k paths that are defined by the TSP+k tour is minimized, and every edge in these paths has a distance that is no longer than any of the resulting k borderline edge lengths.*

3 Application of GTMAS

3.1 The Game Theoretic Aspect of GTMAS

GTMAS has been run to solve the societal clustering problem of the EU countries and Turkey. The normalised 20 features are used as the positions of the cities in a $20-d$ space. The distance between each pair is calculated using these 20 dimensions.

As mentioned above, GTMAS uses the IPD to get the best out off the algorithms (agents) available for the solution of the TSP instance equivalent

of the clustering problem. An example sequence of IPD encounters between the two solver-agents used, after they both have obtained their intermediate solutions, is as follows. Recall, from the PD game, that in each encounter, they either cooperate (C) or compete (D).

Let Solver-Agent 1 (SA1) start and choose to cooperate (C). Let Solver-Agent 2 (SA2) reply by also choosing to cooperate (C). The outcome is a "solution exchange", i.e. each player takes its opponent's solution. The solutions are then evaluated and the payoff table, Table 1, is implemented. Note that in this table **G** indicates a good solution after evaluation, and a **B** a bad one. The player with the good solution is rewarded by gaining an extra unit of CPU time. Note that a unit of CPU time is, here, simply an iteration or the solver-agent's algorithm. The player with the worse solution is penalised by losing two units of CPU time. When both have done well or bad, no reward or punishment is inflicted on either of the players. This is indicated by the letter **x** in the table. Not that this situation is rare.

		SA2	
		B	**G**
SA1	**G**	(1,-2)	**x**
	B	**x**	**(-2,1)**

Table 1. Example Payoff Table for Evaluating and Rewarding Agents

Considereing all combinations for two agents and two playing strategies, four outcomes can result:

1. solution exchange;
2. SA2 takes the solution of SA1 and SA1 does not take the solution of SA2;
3. SA1 takes the solution of SA2 and SA2 does not take the solution of SA1;
4. no solution exchange.

In each case a table similar to Table 1 is drawn and implemented. But, these tables can be combined to give the more compact typical payoff table, Table 2, of GTMAS, in which the player with the good solution, indicated with **G**, is the row player and the player with the worse solution (indicated with a **B**, is the column player, [10] and [12].

The payoffs used are justified as follows. When an agent cooperates it gains one unit (of CPU time or equivalent in terms of iterations it is allowed to do) and loses double that. When it competes it gains two units and loses one (or half of the initial gain). This means, the GTMAS payoff matrix rewards competition. The idea behind supporting competition is to counter the "helping hand" that

cooperation gets from the rules underpinning the construction of GTMAS. It can also be argued that, intuitively at least, too frequent exchanges of solutions will lead to early convergence to local optima. So, competition helps achieve a good coverage of the search space.

		B	
		C	**D**
G	**C**	(1,-2)	(1,-1)
	D	(2,-2)	**(2,-1)**

Table 2. Combined Payoff Table for Evaluating and Rewarding Agents

The equilibrium point for the above payoff matrix is $[D, D]$ with payoffs 2 and -1. It is also a regret-free point. The payoff matrix at the core of GTMAS is different from those commonly found in the literature. These matrices would be drawn immediately after decisions have been taken.

3.2 GTMAS: The general algorithm

The algorithm of GTMAS consists mainly of two procedures: the procedure of the coordinator-agent and the procedure of the solver-agents. Note that solver-agent procedures differ from each other only in the specific algorithm each one of them calls to solve the optimisation problem in hand; here, these algorithms are GA and SA. Otherwise they are the same. To implement competition, CPU time is implicitly used through the number of iterations each solver-agent is allowed to run its algorithm. This is not ideal given the computational load difference between iterations of different algorithms. However, it is easy to implement. Also, because a solution generated by an algorithm can only be taken into account at the end of an iteration, the length of individual iterations is, perhaps, not all that important. A full description of these procedures can be found in [10] and [12].

3.3 Problem data

As already mentioned, the quantitative aspects of the clustering problem are found in EUROSTAT and TURKSTAT data repositories, [4, 14]. Table 3 below contains an excerpt of the much larger table of the scores of each country on each of the 20 criteria listed in Section B.

Table 4 is the corresponding table containing the conversion of these scores into Euclidean distances. Note that values in this table can only be arrived at by

Table 3. Normalised scores of some countries on some of the criteria considered: excerpt from EUROSTATS repository, [4]

	Pop	Marriages	Divorces	MLEB	FemLEB	GDP
Belgium	-47	-18	51	4	2	33
Bulgaria	-61	-20	-36	-6	-5	-64
Czech Rep.	-48	-3	56	0	-1	-17
Denmark	-72	32	40	3	0	37
Germany	304	-6	25	4	1	18
Turkey	256	66	-33	-6	-11	-67

considering the full table from which Table 4 is extracted. The general formula of the Euclidean distance is

$$\sum_{j=1}^{n}(z_{ij} - z_{kj})^2, i, k = 1, .., m, i \neq k$$

where j refers to the criterion and i, k to the countries. The distance between similar countries is put to infinity, here 1000.

Table 4. Euclidean distances computed between some countries on some of the criteria considered

	Belg	Bulga	Czech	Denmar	German	Turkey
Belg	1000	30	19	18	39	31
Bulga	30	1000	20	40	49	37
Czech	19	20	1000	29	43	34
Denmar	18	40	29	1000	44	33
German	39	49	43	44	1000	54
Turkey	31	37	34	33	54	1000

Before carrying out a proper clustering analysis, a pre-emptive exploration of the data by plotting one criterion against an other in a pairwise fashion shows a wealth of results the most interesting of which are as follows. All plots involving religion show Turkey apart from the rest of EU countries. This is a rather obvious result. Less obvious results involve the railways criterion; all plots involving railways show Slovenia apart. This is rather hard to explain until one considers the history of Slovenia and its geographical position. It was, basically a hub for East-West land communication, which meant that an extensive railway network was put in place during the Cold War and prior to it. Slovenia, today finds itself the inheritor of such an infrastructure the extent of which puts it apart from the rest of the EU countries. Similar observations can be made

about certain criteria and certain countries such as Researchers and Finland, and Cinema attendance and Ireland.

3.4 Computational results

The mapped TSP problem is solved 10 times with different numbers of clusters ranging from 1 to 10. In each run, k dummy cities are added to the original cities, k being the number of clusters to be found. The dummy cities have a distance of 5 to the original cities, where 5 is smaller than the smallest distance in the problem and a distance of 1000 to all other dummy cities, where 1000 is larger than any distance. The best number of clusters is found via v-fold cross-validation on the loglikelihood of total point-to-centroid distance.

Table 5 shows the TSP tour lengths and total point-to-centroid distances corresponding to different cluster sizes.

Table 5. V-Fold Cross-Validation Analysis

Nbre of Clusters	Tour Length	Pt-to-Centroid Total Dist	Log (Distance)
1	632.83	692.86	9.44
2	518.47	605.74	9.24
3	493.71	546.28	9.09
4	466.73	530.51	9.05
5	430.21	463.75	8.86
6	407.93	424.77	8.73
7	378.56	384.83	8.59
8	360.72	389.35	8.60
9	333.74	328.88	8.36
10	308.44	342.02	8.42

According to the v-fold cross-validation analysis, the appropriate numbers of clusters are 3 and 7. This is indicated by no or very little change in distance as one moves from a cluster size to the next, in the last column of Table 5. Results for 3- and 7-clustering are reported in Table 6 and Table 7, respectively. Another good solution found for 7-clustering is given in Table 8.

3.5 Comments on the computational results

For 3-clustering, Table 6, Turkey is in the same cluster as Greece, Latvia, Portugal, Ireland, Spain, Italy, United Kingdom (UK), France and Germany.

The measure that determines the members of a given cluster is the overall distance between the countries, representing the overall similarity between

Table 6. 3-Clustering

Cluster 1	Cluster 2	Cluster 3
Turkey	Slovenia	Belgium
Greece		Denmark
Latvia		Finland
Portugal		Poland
Ireland		Slovakia
Spain		Estonia
Italy		Lithuania
UK		Bulgaria
France		Romania
Germany		Czech Republic
		Malta
		Austria
		Sweden
		Netherlands
		Hungary
		Cyprus
		Luxemburg

Table 7. 7-Clustering Alternative Solution 1

Clust 1	Clust 2	Clust 3	Clust 4	Clust 5	Clust 6	Clust 7
Greece	Cyprus	Slovn	Luxm	UK	German	Poland
Latvia	Hung			France		
Portu	Bulga			Spain		
Irlnd	Roman			Italy		
Malta	Czech			Turkey		
Austr	Eston					
Holl	Lithu					
Swden	Slovk					
Belg						
Denmar						
F'land						

them. With this measure, Turkey is found to be closest to Italy, then to Spain and then to Greece. This is an expected result since these countries are rather close culturally, as well as geographically in some cases. After all, they are all Mediterranean countries. Geographical contiguity does not imply similarity. Indeed, the mentioned countries are also in the same cluster as Ireland which is not a Mediterranean country.

Consider the distance table, Table 9, and imagine that Turkey is at the centre of a hypersphere of radius 0. If this radius were increased monotonically, then the first countries to fall into this inflating hypersphere (i.e. the closer countries to Turkey) would be Italy, Spain, Greece, Romania, Poland, Portugal, Latvia, the UK, Bulgaria, France and Lithuania, with distance ranging from 27 to 42. Most of these countries are in the same cluster as Turkey, according to the 3-clustering of Table 6. Thus, the 3-clustering seems to be rather reasonable with respect to the overall distances. Indeed, there are features for which Turkey is different from some of these countries, such as passenger cars and marriages and features for which she is different from any country in the EU, such as religion. Furthermore, the features for which Turkey is closer to each of these countries differ. However, over all features, Turkey is found to be similar to these countries.

Table 8. 7-Clustering Alternative Solution 2

Clust 1	Clust 2	Clust 3	Clust 4	Clust 5	Clust 6	Clust 7
Turkey	Greece	Cyprus	UK	German	Slovn	F'land
	Latvia	Hung	France			Denmar
	Portu	Bulga	Spain			Belg
	Irlnd	Roman	Italy			Luxm
	Malta	Czech				
	Austr	Eston				
	Holl	Lithu				
	Swden	Slovk				
		Poland				

For 7-clustering, there are two alternative good solutions. In one, Table 7, Turkey is with UK, France, Spain and Italy. Greece, Latvia, Portugal and Ireland combine with some other countries from the third group of the 3-clustering and Germany forms a cluster on its own. In the other, Table 8, Turkey is on its own. The UK, France, Spain and Italy are together and Germany is again alone.

These results show that Turkey is in the middle of two groups. The first group is that of Greece, Latvia, Portugal and Ireland which stays close to the

group of Malta, Austria, Netherlands and Sweden, with which they combined in 7-clusterings. The second group is that of Spain, Italy, UK and France, which stay close to Germany on the other side of Turkey. These two groups are placed on different sides of Turkey. In 3-clustering, these groups combined together, and in 7-clusterings, they combine with the other groups they are close to.

The clusterings also give insights about the EU countries themselves and how they cluster together. For instance, Germany is only similar to Italy, Spain, UK and France when the number of clusters is small, however it is separated from every country when the number of clusters are higher. On the other hand, Slovenia is always clustered alone, even in the case of 3-clustering. It is so dissimilar that it forces all the rest of 27 countries into two clusters.

The groups of countries that stay together in each case of clustering are the group of Belgium, Denmark, Finland; the group of Malta, Austria, Netherlands, Sweden; the group of Greece, Latvia, Portugal, Ireland; the group of Italy, Spain, UK, France; and the large group of Hungary, Cyprus, Bulgaria, Romania, Czech Republic, Estonia, Lithuania and Slovakia. These countries are very similar to each other in social terms according to the selected features.

The findings support the idea that Turkey and her people will fit confortably, or at least as well as many existing members, within EU society.

4 Conclusion

The clustering problem has been successfully solved with GTMAS involving two agents one running GA, the other SA. From the different clusters found, Turkey seems to fit well with a number of EU countries. It does so a lot better than some EU countries themselves such as Slovenia.

As stated in [10, 12], GTMAS produces a solution to the given problem by exploiting the synergies between algorithms through cooperation and by selecting the most suited algorithm through competition. Competition is over CPU time. This is achieved via an implementation of the IPD game. In the current setting, with two players, the best results are obtained when both agents compete in the first two stages, especially when GA takes the solution of SA, subsequently. This is unlike the TIT-FOR-TAT strategy which typically starts with a co-operation move.

SA, is less successful than GA over the given runtime. It has not been "eliminated", however, as in this specific environment, it helped GA and the overall performance.

Table 9. Hypersphere Configuration With Turkey at Centre

Country	Order by Dist.	Distance from Turkey
Turkey		0
Italy	1	27
Spain	2	34
Greece	3	37
Romania	4	38
Poland	5	39
Portugal	6	40
Latvia	6	40
United Kingdom	6	40
Bulgaria	7	41
France	7	41
Lithuania	8	42
Slovakia	9	45
Czech Republic	9	45
Austria	9	45
Estonia	10	46
Hungary	10	46
Netherlands	11	47
Germany	12	48
Malta	13	49
Ireland	14	50
Cyprus	15	51
Sweden	16	52
Belgium	17	54
Denmark	18	59
Luxembourg	19	61
Finland	20	66
Slovenia	21	101

References

1. R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.
2. S. Climer and W. Zhang. Rearrangement clustering: Pitfalls, remedies and applications. *Journal of Machine Learning Research*, 7:919–943, 2006.
3. European Union Statistics Institute EUROSTAT. Eurostat pocketbooks, cultural statistics. http://epp.eurostat.ec.europa.eu, 2007.
4. The European Union Statistics Institute EUROSTAT and European Commission. *Europe In Figures, EUROSTAT Yearbook 2005*. EUROSTAT, 2005.
5. J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan, USA, 1975.
6. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:264–323, 1999.
7. J. K. Lenstra. Clustering a data array and the traveling-salesman problem. *Operations Research*, 22(2):413–414, 1974.
8. M. Naznin, P. Juell, K. E. Nygard, and K. Altenburg. A clustering heuristic by effective nearest neighbor selection. In *Midwest Instruction and Computing Symposium*, 2007.
9. The European Union On-Line. Copenhagen criteria. http://europa.eu/scadplus/glossary.
10. Ö. Töreyen. A game-theory based multi-agent system for solving complex combinatorial optimisation problems and a clustering application related to the integration of Turkey into the EU Community, 2008. MSc Thesis in Statistics and Operational Research, Department of Mathematical SCiences, University of Essex.
11. A. Salhi, L.G. Proll, D. Rios Insua, and J. Martin. Experiences with stochastic algorithms for a class of global optimisation problems. *RAIRO Operations Research*, 34(22):183–197, 2000.
12. A. Salhi and Ö. Töreyen. A game theory-based multi-agent system for expensive optimisation problems. In Y. Tenne and C.-K. Goh, editors, *Computational Intelligence in Optimization*, chapter 9, pages 211–232. Springer-Verlag, 2010.
13. StatSoft. Cluster analysis. http://www.statsoft.com/textbook/stcluan.html.
14. The Turkish Statistical Institute TURKSTAT. *Turkey's Statistical Yearbook 2006*. TURKSTAT, Ankara, Turkey, 2007.
15. Secretariat General for EU Affairs / Turkey Negotiation Framework, 2005. http://www.abgs.gov.tr
16. Secretariat General for EU Affairs / Turkey Regular Progress Report for Turkey, 2007. http://www.abgs.gov.tr
17. Brewin, C. Book Review of Turkey and the European Union: Prospects for a Difficult Encounter, 2008.
18. Müftüler-Baç, M. Turkey's Accession to the European Union: The Impact of the EU's Internal Dynamics International Studies Perspectives, 9, pages 201-219, 2008.
19. F-S. Sun and C-H. Tzeng. A Mathematical Model of Similarity and Clustering. In International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 1, pp.460, 2004.
20. I. Davidson and S.S. Ravi. Intractability and Clustering with Constraints. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, OR, 2007. PDF file at http://www.cs.ucdavis.edu/ davidson/Publications/ICML2007.pdf.