

Genetic algorithm applied in Clustering datasets.

James Cunha Werner (wernerjc@sbu.ac.uk)

Terence C. Fogarty (fogarttc@sbu.ac.uk)

*SCISM
South Bank University
103 Borough Road
London SE1 0AA*

Abstract. This paper compares the clustering technique k-means and two different approaches of Genetic Algorithms to a sample dataset, and in EachMovie dataset. The comparison between both techniques shows a better result of GA.

Introduction.

Unsupervised learning is a technique of knowledge extraction from datasets where there is not a correct solution to compare with the estimated output of the model. The best output or model groups the elements with some attribute, building a cluster with minimum total distance with a metric definition.

The final result contains a set of elements that are a base for all elements, with minimise the adopted metrics. All events of one cluster have some attribute in common: kind of movie or book, taste, smell, diagnostic, etc. This information would help to recommend similar thinks with same characteristics to a customer, bringing up new business opportunities.

Preparing data for this kind of problem require some non-objective data acquisition. For example, movie classification would differ for each people between different classes like animation and family, or sometimes horror and comedy, due different points of view, social class and personality. It's difficult to establish a classification for the users too.

To solve this, a large amount of users would be required to establish an average classification for each product, with a mark of like/dislike or a classification. The next step consist in obtain the sets with group all events with a minimum distance, the objective of cluster technique.

This paper describes the techniques k-means and genetic algorithm, and its application in one example for comparative study and the application of genetic algorithms to EachMovie dataset for a proper performance study.

K-means algorithm.

This algorithm is based in an adaptive approach where a random set of cluster base is select from the original dataset, and each element update the nearest element of the base with the average of its attributes (see [1] for a detailed description). To study this algorithm we used an available in Internet software with an example dataset (see [2]).

Genetic algorithm.

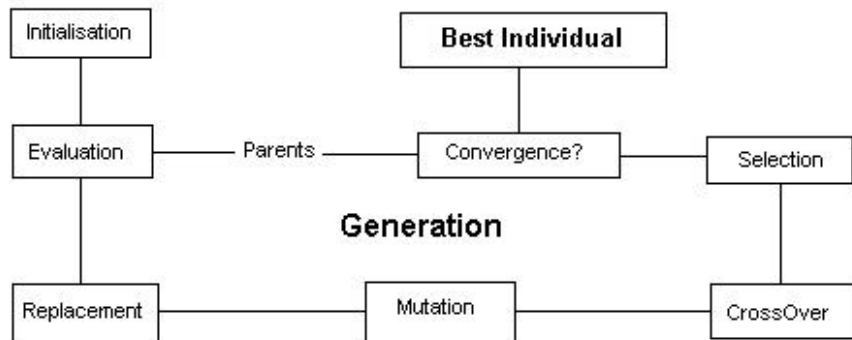
Genetic algorithms (GA) mimics the evolution and improvement of life through reproduction, when each individual contributes with its own genetic information building a new one with fitness to the environment and more surviving chances. These are the bases of genetic algorithms and programming (Holland [3], Goldberg [4] and Koza [5]). Specialized Markov Chains underline the theoretical bases of this algorithm change of states and searching procedures.

Each 'individual' of the generation represents a feasible solution to the problem, coding distinct algorithms/parameters that should be evaluated by a fitness function.

GA operators are mutation (the change of a random position of the chromosome) and crossover (the change of slices of chromosome between parents).

The best individuals are continuously being selected, and crossover and mutation take place. Following few generations, the population converges to the solution that better attend the performance function.

Genetic algorithm concepts and implementation are structured in the following diagram.



Population initialization

Step 1 : Parent selection.

	First population	Objective value $z=f(x,y)$
==>	1100110110101000	3.481746
==>	0101010110110101	3.668023
	1000010100110110	6.261380
	1101011111001100	12.864222

Step 2: Crossover

11010	11111001100	----- \	11010	10100110110
10000	10100110110	----- /	10000	11111001100

Step 3: Mutation

1101010100110110	1000011111001100
1111010100100110	1000011111001100
$z = 8.044$	$z = 6.092$

Step 4: Reinsertion

	second population	Objective value $z=f(x,y)$
==>	1111010100100110	8.044
==>	1000011111001100	6.092
	1000010100110110	6.261380
	1101011111001100	12.864222

Step 5: If the result doesn't converge, go to step 1.

Comparative example.

The data available with the k-means software consist of the name of the element, and 10 attributes. The distance is the Euclidian distance between normalized elements.

GA should apply two different approaches:

1. Coding each component of the cluster in the chromosome: in this case, GA must find each component for each cluster base, resulting a chromosome size with is the number of attributes times the number of cluster in the base.
2. The chromosome contains for each element the cluster it pertains, and the centre is the average of all members. The chromosome size is the number of elements.

The fitness function in both cases is:

$$Fitness = \frac{1.0}{\sqrt{\sum_{cluster\ j} \min \left(\sum_{attribute\ i} (pattern_i - cluster_j)^2 \right)}}$$

mark

The same fitness function could be evaluated for all approaches, and the final value for k-means were 0.138. GA using the number of the cluster for each element (option 2) gave a final fitness of 0.180 and the case of coding each element (option 1) of the base gave 0.189. Both cases the GA result show be more accurate then k-means approach.

Genetic algorithm application with EachMovie dataset.

To apply GA in a clustering problem, we used the EachMovie dataset available in Internet [6], with contains the 1.732.627 votes of 47.636 people for 1.628 movies. The data were filtered remaining the corrects and completes one, from where we normalise and selected the following attributes:

- Person (person.txt) provides optional, unaudited demographic data supplied by each person:
ID: Number -- primary key
Age: Number
Gender: Text -- one of "M", "F"
- Movie (movie.txt) provides descriptive information about each movie:
ID: Number -- primary key
Name: Text
Type: Action, Animation, Art_Foreign, Classic, Comedy, Drama, Family, Horror, Romance, Thriller: Yes/No
- Vote (vote.txt) is the actual rating data:
Person_ID: Number
Movie_ID: Number
Mark or Score: Number -- $0 \leq \text{Score} \leq 1$

We adopted the approach of determine the cluster base for 40 different groups and 3 attributes: sex, age, and type. Type is obtained by the association of 10 bits sequence respectively to the movie type, for example, action is 1 base 2 and 1 base 10, Art_Foreign is 100 base 2 or 4 base 10, etc. Table I shows the more important groups elements.

Male	Female
Cluster: 16 Age: 17 Animation, classic, comedy, drama. # Patterns in Cluster: 118190	Cluster: 22 Age: 11 Animation, Art_Foreign, Classic, Comedy # Patterns in Cluster: 22619
3568 Batman (1989)	441 Independence Day (ID4)
2819 Toy Story	426 Toy Story
2093 The Silence of the Lambs	411 Twister
1978 Seven	330 Mission: Impossible
1957 Ace Ventura: Pet Detective	312 Mr. Holland's Opus
1806 Star Wars	279 Willy Wonka and the Chocolate Factory
1724 Mrs. Doubtfire	256 The Birdcage
1659 Schindler's List	239 The Truth about Cats and Dogs
1636 The Lion King	231 The Nutty Professor
1553 Batman Forever	221 Grumpier Old Men

Male	Female
Cluster: 10 Age: 19 Romance, classic # Patterns in Cluster: 44328 3888 Aladdin 3031 Beauty and the Beast 2810 Goldeneye 2713 Outbreak 2612 Clear and Present Danger 2395 Heat 1820 Addams Family Values 1550 Executive Decision 1322 Four Weddings and a Funeral 1262 The River Wild	Cluster: 28 Age: 20 Romance, Art_Foreign, Animation # Patterns in Cluster: 18553 1625 Aladdin 1471 Beauty and the Beast 903 Clear and Present Danger 880 Four Weddings and a Funeral 875 Outbreak 775 The River Wild 753 Goldeneye 746 Heat 725 Addams Family Values 652 Executive Decision
Cluster: 11 Age: 21 Action, Animation, Art_Foreign # Patterns in Cluster: 252942 6072 Pulp Fiction 5742 True Lies 5122 12 Monkeys 5004 The Fugitive 4585 Mission: Impossible 4315 Twister 4229 Independence Day (ID4) 3956 Broken Arrow 3689 The Rock 3542 Speed	Cluster: 38 Age: 21 Drama, action # Patterns in Cluster: 84599 1198 Independence Day (ID4) 1108 Batman (1989) 1037 Toy Story 1009 The Birdcage 933 Dances With Wolves 893 Twister 880 The Lion King 862 Mission: Impossible 821 The Silence of the Lambs 820 12 Monkeys
Cluster: 14 Age: 28 Drama, comedy, Art_Foreign, Animation # Patterns in Cluster: 156129 6025 Batman (1989) 4413 The Silence of the Lambs 4028 Toy Story 3068 Seven 2948 Dances With Wolves 2876 Star Wars 2610 Mrs. Doubtfire 2600 Schindler's List 2411 Ace Ventura: Pet Detective 2389 Dead Man Walking	Cluster: 17 Age: 24 Thriller, Drama, Animation action # Patterns in Cluster: 19861 1512 Jurassic Park 1436 Father of the Bride Part II 1291 The Shawshank Redemption 945 Terminator 2: Judgment Day 839 Crimson Tide 836 Cliffhanger 696 The Santa Clause 658 Apollo 13 602 Multiplicity 557 101 Dalmatians (1996)
Cluster: 25 Age: 28 Romance, comedy # Patterns in Cluster: 54527 4484 Clear and Present Danger 4281 Aladdin 3509 Beauty and the Beast 3087 Goldeneye 3009 Outbreak 2631 Four Weddings and a Funeral 2524 Heat 2173 Executive Decision 2078 Addams Family Values 1584 Philadelphia	Cluster: 6 Age: 28 Comedy, Art_Foreign, Animation, Action # Patterns in Cluster: 79905 1046 Mission: Impossible 1042 Twister 1009 Independence Day (ID4) 979 Dances With Wolves 974 Pulp Fiction 952 12 Monkeys 922 The Fugitive 919 The Birdcage 917 True Lies 879 Toy Story

Male	Female
Cluster: 2 Age: 32 Action, classic, comedy # Patterns in Cluster: 229666 4445 Pulp Fiction 4368 True Lies 3812 The Fugitive 3704 Dances With Wolves 3222 12 Monkeys 3165 Independence Day (ID4) 3160 Mission: Impossible 3145 Twister 2682 Broken Arrow 2565 Die Hard: With a Vengeance	Cluster: 30 Age: 37 Horror, family, drama, comedy, animation # Patterns in Cluster: 10974 848 Aladdin 842 Beauty and the Beast 669 Clear and Present Danger 565 Outbreak 553 The River Wild 550 Four Weddings and a Funeral 521 Executive Decision 445 Heat 402 Goldeneye 386 Addams Family Values
Cluster: 19 Age: 43 Comedy, classic, animation # Patterns in Cluster: 153114 2699 True Lies 2580 Dances With Wolves 2568 The Fugitive 2268 Pulp Fiction 2182 Independence Day (ID4) 2163 Batman (1989) 2132 Twister 1900 Mission: Impossible 1834 Forrest Gump 1831 12 Monkeys	Cluster: 32 Age: 39 Drama, Art_Foreign, Animation # Patterns in Cluster: 61588 898 Independence Day (ID4) 844 Toy Story 834 Dances With Wolves 825 Twister 772 The Birdcage 764 Batman (1989) 727 Mr. Holland's Opus 686 Mission: Impossible 636 Forrest Gump 632 Sense and Sensibility
Cluster: 7 Age: 49 Drama, comedy, classic, action # Patterns in Cluster: 58880 1742 Batman (1989) 1485 The Lion King 1284 The Silence of the Lambs 1239 Schindler's List 1195 Dances With Wolves 1057 Toy Story 1046 Mr. Holland's Opus 940 Mrs. Doubtfire 937 Dead Man Walking 852 Star Wars	Cluster: 40 Age: 43 Romance, horror, family, drama, art_foreign, action # Patterns in Cluster: 10000 897 Jurassic Park 865 The Shawshank Redemption 792 Father of the Bride Part II 544 Crimson Tide 475 Cliffhanger 423 The Santa Clause 370 Terminator 2: Judgment Day 355 Multiplicity 281 101 Dalmatians (1996) 239 Beverly Hills Cop III
Cluster: 29 Age: 55 Comedy, classics, action # Patterns in Cluster: 56495 1036 Dances With Wolves 1006 The Fugitive 979 True Lies 974 Pulp Fiction 948 Twister 864 Mission: Impossible 844 Independence Day (ID4) 715 Fargo 707 Sense and Sensibility 700 Braveheart	Cluster: 34 Age: 50 Romance, classic, animation, action # Patterns in Cluster: 4797 354 Clear and Present Danger 340 Aladdin 311 The River Wild 294 Executive Decision 291 Beauty and the Beast 289 Four Weddings and a Funeral 241 Outbreak 202 Philadelphia 187 Heat 174 Addams Family Values

The least initial distance of the first generation was 62,321, and the final result after converge was 7,098. The result shows that if I (man with 43 years old) am in Blockbuster lending “Dances with wolves”, should be invited to lend “The fugitive” and “True lies”, which in my opinion is satisfactory.

Conclusion

A class of problems without a training dataset, where the relation between excitation and answer is formally defined, can be addressed by clustering algorithms.

Working with clustering means define a metric distance and the features for what each event will be grouped to obtain the least distance for all set. In EachMovie application, the definition of type of movie as 10 bits adjacent is a feature definition that would affect the solution because, for example, there is a distance between drama (32 base 10 or 100000 base 2) and drama and action (33 base 10 or 100001 base 2) that is not the same (or comparable) as animation (2 base 10) and action (1 base 10). The definition of distance as Euclidian affect this position as well.

The application of genetic algorithms to clustering problems shows better results than kmeans approach, under a controlled example, and the result of the clustering for different movies are reasonable for the author enjoyment.

However, the classification, the data acquisition and the clusters are completely subjective: there is not a correct answer. There is not a function able to classify the same products to any individual, but it would recommend other alternatives and new business opportunities would be created.

References.

1. Bow, S.T; "Pattern recognition applications to large dataset problems"
2. Kyoung-june Min ; <http://vorlon.cwru.edu/~kxm73/> and <http://vorlon.cwru.edu/~kxm73/CW/eces484/eces484.html>
3. HOLLAND, J.H. "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992 reedição 1975.
4. GOLDBERG, D.E. "Genetic Algorithms in Search, Optimisation, and Machine Learning." Reading, Mass.: Addison-Wesley, 1989.
5. KOZA, J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge, Mass.: MIT Press, 1992.
6. Steve Glassman; steve.glassman@compaq.com Compaq Systems Research Center 130 Lytton Avenue Palo Alto, CA 94301