Mining the Eurovision Song Contest

Shaun Tipson
StudentID : 3076294

May 28, 2001

Abstract

The Eurovision Song Contest (ESC) in an event that was created in 1956 by the European Broadcasting Union (EBU), and has been held annually ever since. Each country participating in the competition performs an original song, and votes are then given based (presumably) on the merit of each song.

The song that attracts the most votes is declared the winner of the ESC for that year, and the Contest is generally hosted by the winners' country the following year.

This paper looks at the voting patterns of countries participating in the ESC. Each voting transaction (a *voting* country giving a number of *votes* to a *receiving* country) is considered atomically. This set of atomic transactions has been mined using the Weka application [6].

The overall hypothesis is that the quality of a song is *not* the only (or even dominant) characteristic that is relevant when predicting how many votes it will receive. Many commentators have claimed that geographical location, political alliances and even where the song is performed in the contest, may instead be used as better predictors. The aim of this paper is to see whether this is true or not.

1 Introduction

The Eurovision Song Contest (ESC) began in 1956 with only 8 countries competing. Various factors - including the breakup of the Soviet Bloc and the greater range and sophistication of television broadcasts - have caused the number of competitors to gradually climb. This number reached a height of 25 in 1998, and has since been limited to 23 countries per year.

The remainder of the countries that wish to compete take the place of the worst performing songs from the previous year. For example at ESC 2001 held in Copenhagen, Denmark, the bottom 8 countries were relegated and are thus ineligible to compete in ESC 2002 (to be held in Tallinn, Estonia).

The ESC voting data is suitable for data mining due to the atomic nature of each vote. The construction of each voting tuple is discussed further below (section 2.3). Although statistical analysis of the data has been performed by other people (see for example [3]), it seems that no-one has looked at it from a data mining point of view.

This means that the focus has generally been on a yearly rather than vote-by-vote basis (how many points to win the competition, what advantage is there in hosting it etc.). The advantage of the data mining approach is that variations introduced by the quality of a song in any given year should be "drowned out" by the rest of the data set. In other words song quality will only be reflected in the data if the quality of a country's entry is consistent[1] from year to year.

The real interest in this data set comes in two parts:

Allegiances: To what extent can a voting country separate itself from traditional allegiances and enmities when voting;

Prediction: To have some basis for the claim that Norway always votes for Sweden or that singing last will get you more votes. This information could also enable you to win the ESC Drinking Game! [5]

2 The Data

The data was taken from an ESC Statistics web-site [3]. I wrote a Perl script to parse the HTML pages and extract the voting information for the years 1974-2000 (inclusive). The script then created an output file in which it stored comma separated tuples with one voting transaction per line. This file could the easily be converted to ARFF format - the file format accepted by Weka (see Appendix A).

The data for 1966-1973 was available but ignored due to the different scoring system employed in those years. Data prior to 1966 was unavailable but probably unusable anyway.

Two data sets were eventually examined: one with **11664** voting transactions and the other with **5641**. The reason for two different data sets is explained below in section 2.4.

[1] Good or bad!

2.1 History and Eligibility

The set of countries that competed in the period 1974-2000 is presented in the following table (the country names have been extracted from the ARFF file).

AUSTRIA	BELGIUM	BOSNIA	CROATIA
CYPRUS	DENMARK	ESTONIA	FINLAND
FRANCE	FYRMACEDONIA	GERMANY	GREECE
HUNGARY	ICELAND	IRELAND	ISRAEL
ITALY	LATVIA	LITHUANIA	LUXEMBOURG
MALTA	MONACO	NETHERLANDS	NORWAY
POLAND	PORTUGAL	ROMANIA	RUSSIA
SLOVAKIA	SLOVENIA	SPAIN	SWEDEN
SWITZERLAND	TURKEY	UNITEDKINGDOM	YUGOSLAVIA
MAROCCO			

There are a number of issues that come out of this list:

- Some countries have only existed or participated in the Contest for a short amount of time (for example Russia).
- By the same token some countries have ceased to participate or exist (for example Italy and Yugoslavia).

In both of these cases the countries will only be represented in a limited number of voting tuples. This is important when considering the *support* of any association rules (see section 4.2).

- There is also an issue of data integrity, since there is obviously no country called "Maroco". Since I couldn't decide whether is should have been Monaco or Morocco (both countries are EBU members), I left the data as is.

2.2 The Scoring System

Eurovision votes are given according to the following procedure:

- Each participating country gives out 12,10,8,7,6,5,4,3,2 and 1 vote(s) to songs that have been performed in the Contest.
- Votes are given one country at a time in order of performance (for example if Germany performs first then it also gives out all of its votes first).
- No country may give any votes to itself.
- The country with the most votes after every participant has voted is the winner.

Up to the mid 90s, votes were generally given by a jury of experts from each country. The voting process has now been changed so that people watching the contest can call up and vote for their favourite song.

This transition in voting has been ignored in the data analysis (since it would be too complex to include). It is arguable, though, that voting patterns changed once voting became popular rather than expert.

2.3 The Original Data

Each atomic transaction is represented in the data file by the following seven-attribute tuple:

$$\{voting\_country,\ receiving\_country,\ num\_votes,\ year,\\ singing\_order,\ voting\_region,\ receiving\_region\}$$

The attributes had the following types:

Country One of the set of countries.

Num Votes An integer of the set {1,2,3,4,5,6,7,8,10,12}

Year An integer in the range [74..100]

Singing Order An integer in the range [1..25][2]

Region One of the set of regions (see below).

This information was extracted from [3]'s HTML pages to produce **5641** transactions. Two different regional groupings were performed on the data in order to simplify analysis.

Regions were created using an equi-width rather than equi-depth philosophy (ie there are approximately the same number of countries in each group but *not* approximately the same number of voting transactions, see [1]).

An example data file is presented in Appendix A.

2.3.1 Countries by Political Region

The splits are fairly obvious apart from Israel (placed in the English category for strategic regions) and Marocco (the non-existent country!).

Scandinavian	Baltic	English	Continental
Denmark	Latvia	United Kingdom	France
Sweden	Lithuania	Ireland	Netherlands
Iceland	Estonia	Malta	Belgium
Norway		Israel	Luxembourg
Finland			Monaco

Eastern	Balkan	Central	Mediterranean
Bulgaria	Bosnia	Switzerland	Greece
Poland	Croatia	Germany	Italy
Russia	Slovenia	Austria	Turkey
Romania	Yugoslavia		Cyprus
Slovakia	FYR Macedonia		Spain
Hungary			Portugal
			Marocco

[2] Where in the contest the country performed (ie 1st, 2nd...).

2.3.2 Countries by Geographical Region

All fairly obvious except the central European countries (which have all been placed in the West category). These classifications also reflect political and economic differences (which are themselves the result of geography).

North	East	South	West
Denmark	Bulgaria	Malta	United Kingdom
Sweden	Poland	Israel	Ireland
Iceland	Russia	Greece	Switzerland
Norway	Romania	Italy	Germany
Finland	Slovakia	Turkey	Austria
Latvia	Hungary	Cyprus	France
Lithuania	Bosnia	Spain	Netherlands
Estonia	Croatia	Portugal	Belgium
	Slovenia	Marocco	Luxembourg
	Yugoslavia		Monaco
	FYR Macedonia		

2.4 Variations on the Original Data

In the course of analysing the data a number of modifications were experimented with in order to make the results clearer:

Normalisation The mainly applied to the singing_order attribute, since the number of competing countries changed from year to year.

Instead of being expressed as a number of the set $[1..25]$[3], singing_order was changed to an integer in the range 0 to 100. On this scale, 50 means that a country performed at the halfway point for a given year, and 100 means that their performance was last.

Discretisation Integer sets were changed to classes to allow easier classification and association.

- **Countries** were mapped to a given region. This has already been discussed above (sections 2.3.2 and 2.3.1).
- **Votes** were mapped to the set {*low, medium, high*}. The partition was done in equal widths.
- **Singing Orders** were mapped to the set {*early, mid-early, mid-late, late*}. The partition was done in equal widths.

Inclusion If there are 23 countries competing and each country gives out 10 lots of votes, this means that 12 countries receive no votes at all for a given round.

Including the "zero" votes can be useful since it can indicate who countries' *don't* like to vote for. Including these votes enlarged the data set to **11664** transactions, which turned out to be useful for classification but not association.

[3] Where the higher numbers would not be used at all in most years.

3 Making Predictions About the Data

Before analysing the data it is worth stating a few of the rules that I thought would be supported:

Regions Vote For Each Other It seems like Norway always votes for Sweden and Estonia always votes for Latvia. These "voting blocks" are what make the ESC voting process so entertaining.

Allegiances Some countries love each other. Here are a few samples (disregarding regional alliances):

- Greece and Cyprus
- Germany and Turkey
- Spain and Portugal

Enmities Similarly, some countries hate each other:

- Ireland and UK (but not vice versa)
- Turkey and {Cyprus, Greece}
- Poland and Germany
- France and UK

Language Groups Countries will like other countries who speak the same or similar language

Singing Last is Best By the time you get to the end of 23 songs it is probably hard to remember what the first one sounded like! The most recent impression should get the most votes.

Cultural Cringe and Imperialism Eastern and poorer countries tend to admire and copy their more prosperous Western neighbours.

3.1 Wogan's Rules

In addition to these general rules, I thought that I would include a few opinions expressed by BBC broadcaster Terry Wogan, who commentated for ESC 2001.

- **Iceland and Denmark:**An extension of the regional voting rule.
- **Generous Sweden:** Sweden will tend to vote for less privileged countries.
- **Ireland likes Abba:** The Irish will always give a lot of votes to Scandinavia.
- **Israel votes politically:** The Israelis will vote on the basis of who their friends are.
- **Eccentric French:** The French votes are generally unpredictable and arbitrary.

4 Analysis

Figure 1 has been included to show that certain countries will be involved in many more voting transactions then others. For example Spain (SPA) has been involved in every ESC in the data, whereas Lithuania (LIT) has appeared only twice (in 94 and 99).

Any rules involving countries with a limited data set (such as Lithuania) may be difficult to support without more data.

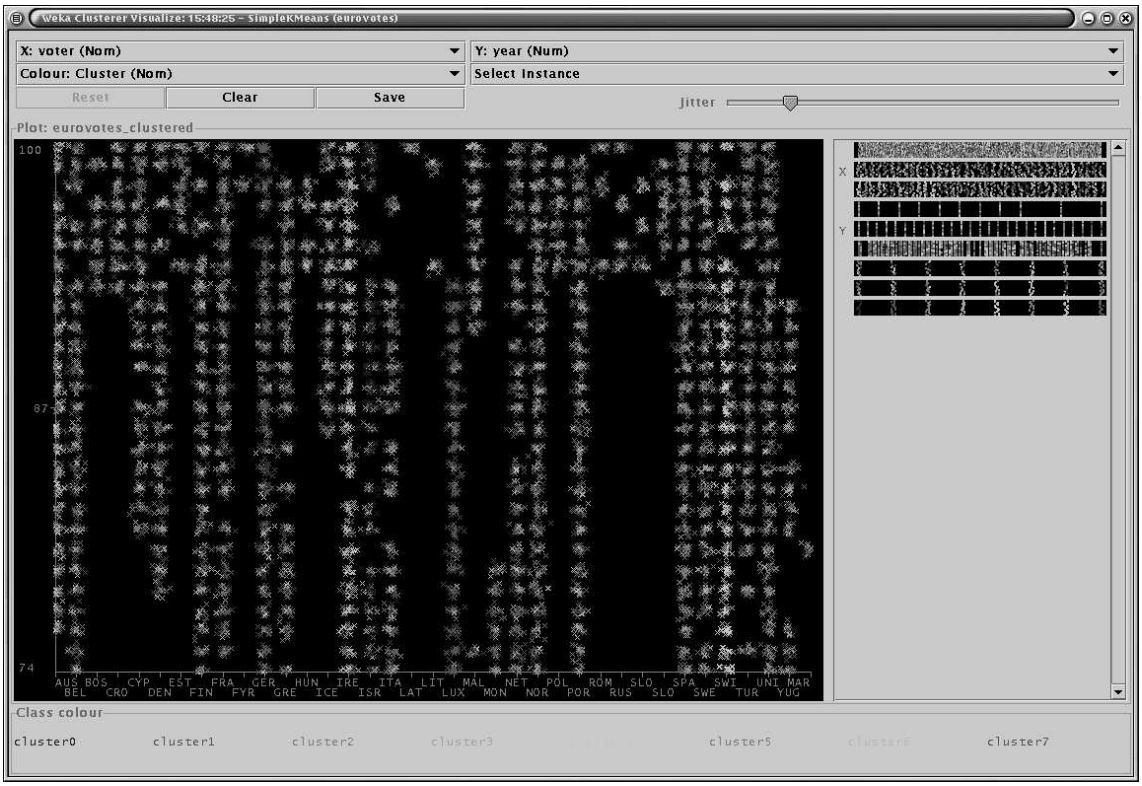

Figure 1: Country Participation by Year

4.1 Classification

Classification was used to see if there were any relationships between voter, receiver and number of votes. The other four attributes were ignored in order to keep the tree size to a minimum.

The main classifier used was `weka.classifiers.j48.J48`, which implements the C4.5 decision tree learner. The algorithm produces a decision tree along with classification statistics and a confusion matrix.

4.1.1 Branches Involving Countries

- **Attributes:** {*voter_country, receiver_country, num_votes*}
  Rule: `if` *voter_country* `and` *num_votes* `then` *receiver_country*
  Number of Leaves: 284
  Size of Tree: 532
  Correctly Classified Instances: 11.9%
  Total Number of Instances: 11640

Although the classification rate is not high, there are still some useful sub-trees that can be extracted from the data. Figures 2 to 6 each display a sub-tree of the larger tree generated by Weka.

In each of the trees there are three levels: the number of votes (high or zero), the voting country and the receiving country. The leaves of the tree (and therefore the class we are trying to predict) are receiving countries.

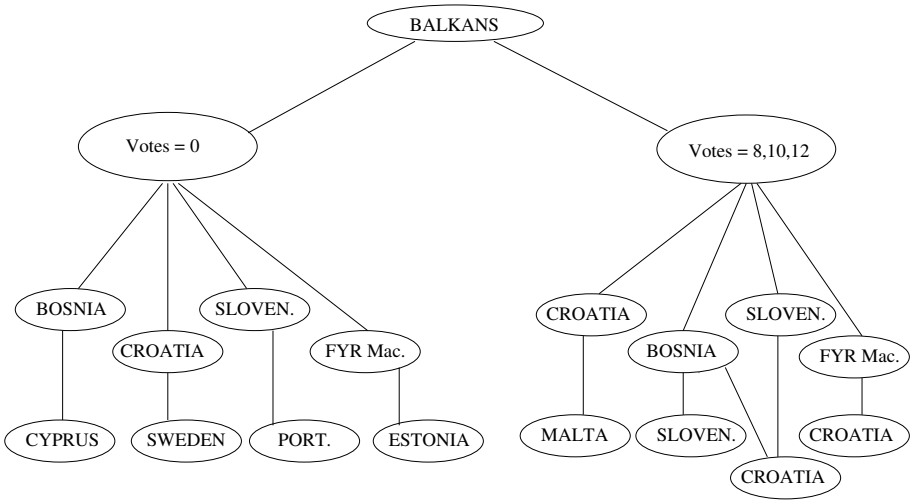

Figure 2: Balkan Voting Strategies

The sub-trees have all been selected on the basis of interestingness and there relation to the rules expressed in section 3. Conclusions can be drawn from each sub-tree:

Figure 2 The Balkans really like to stick together. There seems to be a strong chance that high votes will go to a neighbour.

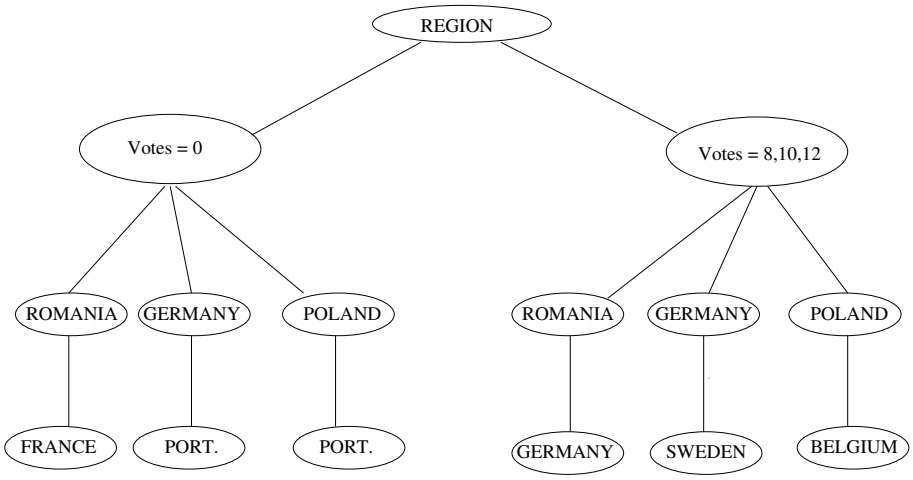

Figure 3: Germany's Friends and Enemies

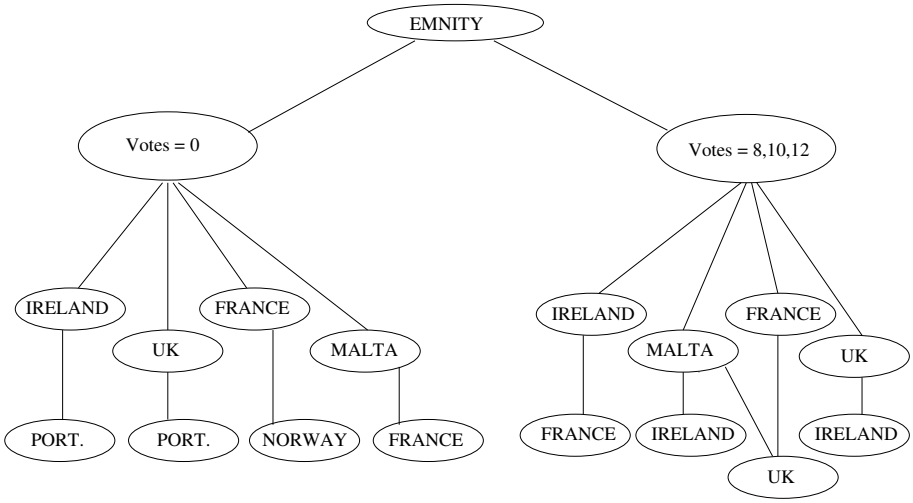

Figure 4: Traditional Rivals and Allies

Figure 3 The regional loyalty is not quite as strong here, although it seems likely that the zero votes will not go to the neighbours.

Figure 4 This tree seems to indicate that the UK doesn't mind giving votes to Ireland but the opposite is *not* true. On the other hand the UK is likely to receive high votes from France (probably a "no hard feelings" vote) and Malta (probably language-based).

Figure 5 The Scandinavian countries are another region that likes to stick together. High votes go to neighbours and zero votes go to far away (mostly Spain and Portugal).

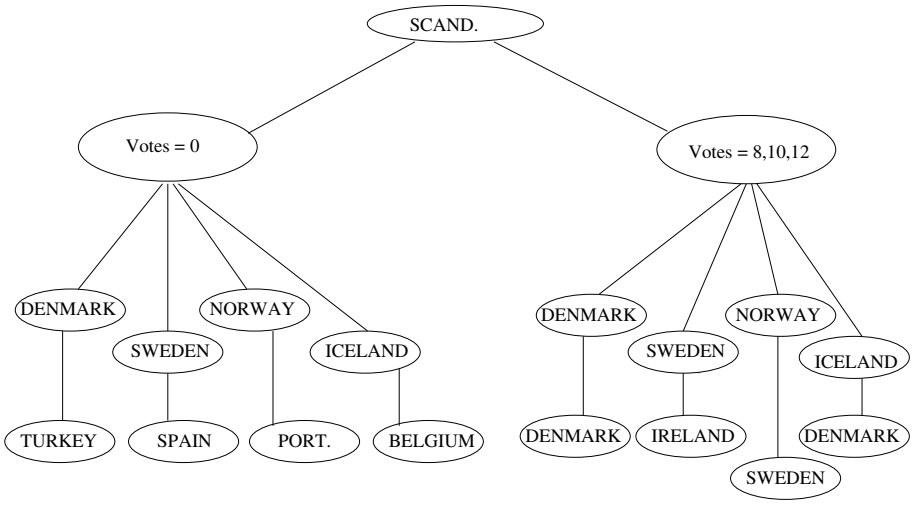

Figure 5: Scandinavia's Regional Voting

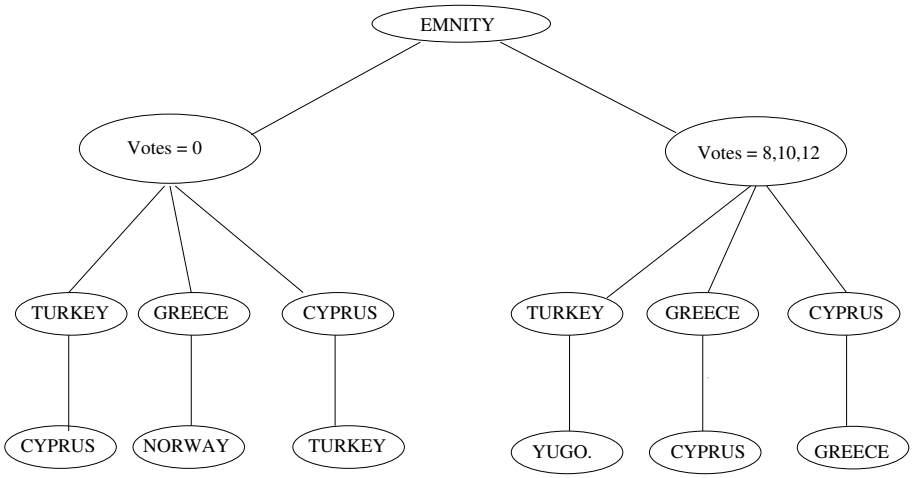

Figure 6: The Relationships Between Turkey, Greece and Cyprus

Figure 6 The three way relationship between Greece, Turkey and Cyprus is the strongest of them all. Greece and Cyprus will vote for each other but not Turkey. The tree indicates that the opposite is also true. The political reason for these voting patterns is obvious, however the strength of the correlation is surprising.

There already seems to be a tendency for some regions to prefer to vote internally. This pattern should be expected to be even more obvious when classification is done by region.

4.1.2 Branches Involving Regions

- **Attributes:** {*voter_region, receiver_region , num_votes*}
  Rule: `if` *voter_region* `and` *num_votes* `then` *receiver_region*
  Number of Leaves: 41
  Size of Tree: 69
  Correctly Classified Instances: 27.1%
  Total Number of Instances: 11664

The classification algorithm predicted that the region most likely to receive zero votes was the Mediterranean. The only region that didn't follow this rule was the Mediterranean itself. Med countries were most likely to give zero votes to Scandinavia.

This is interesting because there were approximately the same number of voting tuples containing Mediterranean countries as there were of any other class. This implies:

1. Mediterranean countries produce very bad songs;
2. Mediterranean musical taste is different from the rest of Europe or;
3. Mediterranean countries haven't been in the Contest enough.

I would suggest it is a combination of these factors.

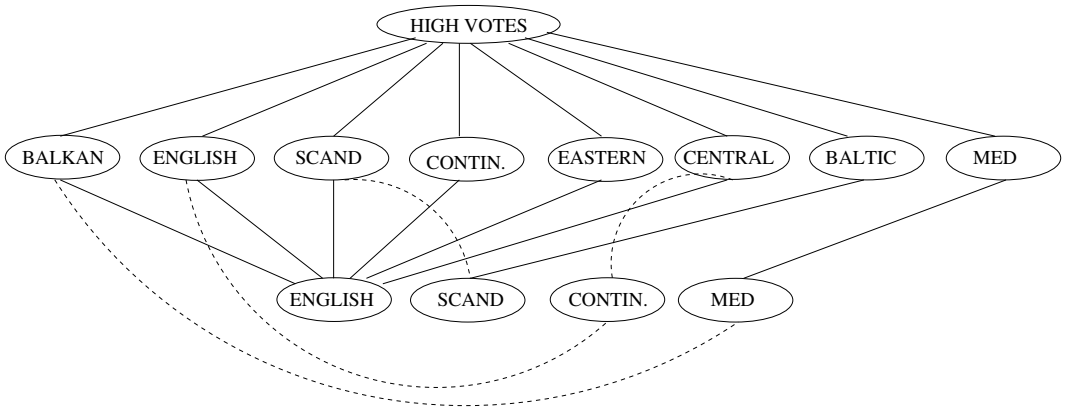

Figure 7: Regional Voting

Figure 7 displays the more general voting preferences of regions (the second preference - if it was clear from the data - is displayed by a dotted line). Surprisingly, regional alliances are not particularly clear.

The reasons for this are two-fold:

1. English speaking countries have won the competition a disproportionate number of times (especially Ireland with 6 wins). As a consequence they have attracted more and greater votes.
2. The tree pruning process has mis-classified a large number of votes from the smaller regions (in order to make the tree simpler). For example, the

small number of votes cast by Baltic countries has been overwhelmed by the large number of Scandinavian votes.
This problem arises again in relation to association rules (section 4.2).

Despite these problems, there a still a couple of interesting features from the regional tree. Most notably, there seems to be northern[4] and southern[5] European voting blocs.

4.1.3 Other Classification Strategies

Two alternative classification strategies were also tried:

1. Using the receiving country to predict the voting country (in the previous section is was the other way around;
2. Using the PART algorithm to classify the data.

The first case proved to be useful since it confirmed the results from the first run and produced evidence that supported some of the initial predictions (section 3).

- The Scandinavian countries and Balkan countries were still sticking together;
- Votes from France were likely to go to either Luxembourg or Belgium (presumably on the basis of language);
- Germany was likely to give votes to Turkey;
- Spain and Portugal were likely to give each other a 1,2 or 3 (but no higher) votes.

The PART algorithm turned out to be less useful since it correctly classified only 6% of the data. However, the rules that were produced seem to confirm what had already been extracted by the C4.5 algorithm.

4.2 Association Rules

- **Attributes:** All except *year*
  Algorithm: Apriori
  Large Itemsets: 100 (approx.)
  Minimum Confidence: 0.3 (approx.)
  Total Number of Instances: 11664

Association rules were extracted from the two data sets. The smaller data set (without the votes) was generally more useful since the support of rules not involving zero votes was increased. In analysing the data a special vote category of "zero" was introduced into the data.

The following association rules were produced:

[4] {*English, Scandinavian, Baltic, Continental*}
[5] {*Mediterranean, Balkan*}

Num_votes and Singing_order The overall pattern seemed to be that singing later was better (as predicted). The three sets of rules are from the big, small and big data sets respectively:

- {votes = high, medium} $\longrightarrow$ {order = mid_late, late} (conf = 0.29)
  {order = early, mid-early} $\longrightarrow$ {votes = low} (conf = 0.27)
- {order = mid-late, late} $\longrightarrow$ {votes = high} (conf = 0.3) **BUT**
  {order = early, mid-early} $\longrightarrow$ {votes = high} (conf = 0.27)
- {order = 1} $\longrightarrow$ {votes = zero} (conf = 0.55)
  {order = 2} $\longrightarrow$ {votes = zero} (conf = 0.59)
  {order = 5} $\longrightarrow$ {votes = zero} (conf = 0.49)
  {order = 6} $\longrightarrow$ {votes = zero} (conf = 0.56)

Num_votes and Regions The countries in the English region have a much lower proportion of low votes (about 34%) then any other region (about 42% for everyone else) in the data set without zero votes. This means that when votes are given to English countries they are generally high.

Apart from the English, with the large number of regions it was difficult to extract useful association rules. As a result the countries were re-partitioned into {NORTH, EAST, SOUTH, WEST} and the Apriori algorithm was run again:

- {rec_region = MED} $\longrightarrow$ {votes = low} (conf = 0.45)
  {votes = high, medium} $\longrightarrow$ {rec_region = ENGLISH} (conf = 0.25)
- {votes = low} $\longrightarrow$ {rec_region = WEST} (conf = 0.44)
  {votes = medium} $\longrightarrow$ {rec_region = WEST} (conf = 0.44)
  {votes = high} $\longrightarrow$ {rec_region = WEST} (conf = 0.49)
- {votes = high} and {vote_region = WEST} $\longrightarrow$ {rec_region = WEST} (conf = 0.55)
  {vote_region = WEST} $\longrightarrow$ {rec_region = WEST} (conf = 0.49)
- {votes = med} and {vote_region = SOUTH} $\longrightarrow$ {rec_region = WEST} (conf = 0.49)
  {vote_region = SOUTH} and {rec_region = WEST}$\longrightarrow$ {votes = med} (conf = 0.49)

Other Interesting Rules There were several other interesting rules in the list generated by Weka:

- {order = mid-late} and {vote_region = WEST} $\longrightarrow$ {rec_region = WEST} (conf = 0.55)
- {order = early} $\longrightarrow$ {votes = medium} (conf = 0.41)
  {order = mid-early} $\longrightarrow$ {votes = medium} (conf = 0.39)

4.3 Clusters

For completeness (and because Weka makes it easy), the data was also ran through through the Simple K-Means clustering algorithm. Weka was asked to find 8 and 20 clusters for the {*voter_region,receiver_region*} set.

There doesn't seem to be any patterns in the clusters except for a general separation into northern and southern voting blocs. The 8 cluster results are presented in Appendix B.

5 Conclusions

The results seem to be a bit patchy. It seems like some regions (Scandinavia, Balkans) are much more tightly knit then others (Med, Continental). Despite this fact, there is definitely enough support in the analysis for the proposition that regions tend to stick together.

It can also be seen that relationships play an important role in the voting patterns of some countries. There is some degree of consistency whenever the UK, Ireland, France, Turkey, Greece, Cyprus and Germany vote. On the other hand, {Spain and Portugal} and {France and the UK} don't behave as you would expect.

There is also support for the proposition that singing later is better. The association rules tend to indicate that the high votes go to the late performers and the zero votes to the early ones.

The final strong support is for the notion that language groups stick together. The classification process revealed groups for English, French and Scandinavian speaking countries.

The rest of the predictions are not as easy to prove (especially Terry Wogan's). Votes tend to flow from West to East and South to North but that might just be the result of Western countries having participated more.

As far as support for Wogan's rules is concerned:

- Iceland and Denmark do vote for each other, but this is just part of the greater Scandinavian alliance;
- Sweden doesn't seem to vote for poor countries and Ireland doesn't seem to vote for the North;
- It is impossible to tell whether Israel has voted politically. The only solid conclusion from their voting patterns is that they tend to vote for wealthy Western European countries;
- The French seem to be as predictable as anyone else. This rule is probably just the product of the fact that Wogan is English.

Overall, a country is better off if it is singing towards the end and it has lots of neighbours in the Contest. This was shown to be true this year in ESC 2001 where the winner (Estonia) performed 20th out of 23, and was supported by the Baltic states and Scandinavia.

5.1 Further Approaches

In addition to the seven attributes used above (and *year* was hardly touched on), there are a few other things that might be worth looking at:

- The gender of the performer;
- The language that the song was performed in;
- Defining neighbour as "sharing a border";
- Defining geographical relationships by "number of kilometres from border to border".

Since extracting this information would be a time consuming process, it is doubtful whether the effort would be worthwhile.

Appendix A: Sample Data File

The following is an extract from one of the ARFF files used in the data analysis process. This file was an early version and makes no distinction between the CENTRAL and CONTINENTAL regional classes.

```
%
% Data file for Eurovision votes
%
@relation eurovotes
@attribute voter {AUSTRIA, BELGIUM, BOSNIA, CROATIA, CYPRUS, DENMARK,
ESTONIA, FINLAND, FRANCE, FYRMACEDONIA, GERMANY, GREECE, HUNGARY,
ICELAND, IRELAND, ISRAEL, ITALY, LATVIA, LITHUANIA, LUXEMBOURG,
MALTA, MONACO, NETHERLANDS, NORWAY, POLAND, PORTUGAL, ROMANIA,
RUSSIA, SLOVAKIA, SLOVENIA, SPAIN, SWEDEN, SWITZERLAND, TURKEY,
UNITEDKINGDOM, YUGOSLAVIA, MAROCCO}
@attribute receiver {AUSTRIA, BELGIUM, BOSNIA, CROATIA, CYPRUS,
DENMARK, ESTONIA, FINLAND, FRANCE, FYRMACEDONIA, GERMANY, GREECE,
HUNGARY, ICELAND, IRELAND, ISRAEL, ITALY, LATVIA, LITHUANIA,
LUXEMBOURG, MALTA, MONACO, NETHERLANDS, NORWAY, POLAND, PORTUGAL,
ROMANIA, RUSSIA, SLOVAKIA, SLOVENIA, SPAIN, SWEDEN, SWITZERLAND,
TURKEY, UNITEDKINGDOM, YUGOSLAVIA, MAROCCO}
@attribute votes {1,2,3,4,5,6,7,8,10,12}
@attribute year integer
@attribute singing_order {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,
17,18,19,20,21,22,23,24,25}
@attribute voter_region {SCAND,BALTIC,ENGLISH,EASTERN,BALKAN,CENTRAL,MED}
@attribute receiver_region {SCAND,BALTIC,ENGLISH,EASTERN,BALKAN,CENTRAL,MED}

@DATA
FINLAND,UNITEDKINGDOM,1,74,2,SCAND,ENGLISH
FINLAND,YUGOSLAVIA,1,74,7,SCAND,BALKAN
FINLAND,SWEDEN,5,74,8,SCAND,SCAND
FINLAND,NETHERLANDS,1,74,12,SCAND,CENTRAL
FINLAND,ITALY,2,74,17,SCAND,MED
UNITEDKINGDOM,FINLAND,1,74,1,ENGLISH,SCAND
UNITEDKINGDOM,ISRAEL,2,74,6,ENGLISH,MED
```

Appendix B: Eight Cluster Results

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 8 -S 10
Relation:     eurovotes
Instances:    11664
Attributes:   2
              voter_region
              receiver_region
Test mode:    evaluate on training data

=== Clustering model (full training set) ===


kMeans
======

Cluster centroids:

Cluster 0
         ENGLISH MED
Cluster 1
         CONTINENTAL MED
Cluster 2
         MED CONTINENTAL
Cluster 3
         MED BALKAN
Cluster 4
         CENTRAL MED
Cluster 5
         MED SCAND
Cluster 6
         SCAND CENTRAL
Cluster 7
         SCAND SCAND
```

References

[1] J. Han and M. Kambar,*Data Mining: Concepts and Techniques*, Morgan Kauffman, 2001.

[2] The European Broadcasting Union (http://www.ebu.ch)

[3] Eurovision Song Contest Statistics (http://www.kolumbus.fi/jarpen/)

[4] Official Eurovision Song Contest Website (http://www.songcontest.com)

[5] The Eurovision Song Contest Drinking Game (http://www.cowgate.demon.co.uk/mit/drinking/eurovision.html)

[6] The Weka Set of Data Mining Tools (http://cs.waikato.ac.nz)