Paper 122-28

Shopping for Voters: Using Association Rules to Discover Relationships in Election Survey Data

Mary MacDougall, SPS Software Services, Inc., Cleveland, OH

ABSTRACT

Association analysis is an undirected data mining technique that is especially helpful for exploring new data sets because it provides results that are easy to understand and explain. While it is commonly used for finding relationships between events in retail settings, such as the purchase of beer and diapers, association rules can also highlight connections between intangible things like a person's age, his political views, and his TV viewing habits. This paper describes a case study in which the Association node in SAS Enterprise Miner™ was used to examine relationships between the demographic characteristics, political party affiliation and media influences of respondents in a survey of U. S. voters.

INTRODUCTION

The American National Election Studies (NES) are surveys conducted by the Center for Political Studies of the Institute for Social Research at the University of Michigan immediately before and after U. S. national elections (Burns, 2001). In this case study using Enterprise Miner software, a team made up of five students in a data mining course at Cleveland State University chose to pose as political consultants analyzing the 2000 National Election Study in order to 1) discover something interesting about the data and 2) gain experience with the data mining process.

The NES data have several features that make it a good choice for a data mining case study. The study methodology is well documented, and the data are available in the form of a SAS data set. Each survey question is represented by a numeric variable and a corresponding format, which provides descriptions of the coded response values. Invalid or non-applicable responses are coded with special values such as 99.

Although the data set size of 1800 observations is small, the large number of interview questions covers a wealth of detail about each respondent's demographic and social background, and political views.

EXPLORATION

The second step in SEMMA, the data mining methodology by SAS Institute, is Explore – Search speculatively for unanticipated trends and anomalies so as to gain understanding and ideas. (SAS Institute, 2001). Exploration was an important step toward developing meaningful predictive models in the study, especially because the project team in this study did not have a background in political science.

The team initially tried to use Clustering to find groups of people with similar patterns of responses. Although the Clustering node produced intriguing charts showing distinct groupings, it was difficult to interpret how the clusters were defined.

We also attempted to use decision trees to understand the data, but the tree results did not make intuitive sense either. They seemed to indicate that whether or not a respondent recognized the Prime Minister of England and how often he or she watched "Jeopardy" on television seemed to be critical, but we did not know what to do with this information.

We finally selected Association Rules, otherwise known as Market Basket Analysis, as an exploration tool to help us discover the underlying patterns in the survey responses.

DATA TRANSFORMATION

Although the NES study data set was relatively clean, compared to the typical data that might be extracted from a corporate transaction system, the following steps were still necessary to transform the interview responses into inputs that could be used by the Association node:

1. Exclude the 252 interviews that did not have a post-election follow-up interview. The remaining 1,555 records had complete information, including variables showing whether the respondent voted and for whom.
2. Derive new variables. For example, the answers to a series of twelve questions about the relative trustworthiness, intelligence and work ethic of people of different races were used to derive a new variable called RACE_BIAS. RACE_BIAS was set to 1 if the respondent assumed different character traits based on race, and 0 if the respondent did not differentiate based on race.

   REGION was derived from grouping the respondents' home states into Northeast, South, Midwest and West. STATE was then dropped from the analysis data set to avoid redundant rules.
3. Reduce the number of different responses in order to make the model results easier to interpret. For questions that had a range of possible responses such as: agree strongly, agree somewhat, disagree somewhat, disagree strongly, responses were converted into binary values of 1 or 0, representing "For" or "Against".
4. Group variables into categories. With over 1,800 interview questions to choose from, it was difficult to select variables for our models. We split the variables into the following categories: Demographics, Party Affiliation, Policy Issues, Political Information Sources, and Political Involvement, and assigned each category's list of variables to a macro variable for easy reference.
5. Transpose responses into rows. Each record in the source data contained responses from one interview. In order to use the Association Node, it was necessary to transpose the response values into multiple rows, with one record for each item to be considered in the analysis.

   The following macro was used to transpose the binary responses into an output row when the response was positive.

```
%macro binary (v, a);
  if &v then do;
     item = &a;
     output;
  end;
  else item = ' ';
%mend binary;
```

   Another short macro applied a custom format to the variables with multiple responses and transposed them into activities.

```
%macro fmt (v,a);
  item = put(&v,&a);
  output;
%mend fmt;
```

The table in Figure 1 shows a sample of records from one of the transposed data sets.

RESPONDENT	ITEM
1510	FEMALE
1510	WHITE
1510	AGE 65+
1510	SOUTH
1510	DEMOCRAT
1510	NATL NEWS
1510	EARLY LOCAL NEWS
1510	NEWSPAPER
1511	FEMALE
1511	WHITE
1511	AGE 18 - 29
1511	WEST

Figure 1.

ASSOCIATION ANALYSIS

For our first analysis, we included items from the demographics, political information sources, and party affiliation. The second analysis we ran included items from the demographics, party affiliation and policy issues groups. Compared to the results generated by Clustering, the association rules produced by the Association Node in Enterprise Miner were refreshingly easy to understand.

FREQUENCIES

Although, by default, Enterprise Miner displays the list of rules generated, a first look at the results should start with the Frequencies tab. The table in Figure 2 lists the frequencies for the Policy Issues analysis. Note that the items that describe demographic characteristics are mixed in with the items indicating the respondent's position on policy issues. The Association Node does not distinguish between different types of items.

COUNT	ITEM
1245	USE SURPLUS FOR SOC. SEC.
1236	WHITE
1165	RACE BIASED
1109	FOR GAYS IN MILITARY
1078	FOR DEATH PENALTY
1012	NO CHILDREN AT HOME
968	COLLEGE
951	USE SURPLUS FOR TAX CUTS
908	FOR GUN CONTROL
881	FEMALE
776	FOR ENVIR. PROTECTION
674	MALE
665	MIDDLE CLASS
664	WORKING CLASS
645	FOR SCHOOL VOUCHERS
636	FOR ABORTION RIGHTS
627	INDEPENDENT
587	HS OR LESS
548	SOUTH
543	CHILDREN AT HOME
526	DEMOCRAT
524	AGE 45 - 64
503	AGE 30 - 44
458	URBAN
435	RURAL
402	REPUBLICAN
392	MIDWEST
373	NONVOTER
345	WEST
342	SMALL TOWN
320	SUBURBAN
291	AGE 65 +
270	NE
257	FOR WELFARE PROGRAMS
229	AGE 18 ╤ 29
226	UPPER MID CLASS
162	BLACK
69	HISPANIC

Figure 2.

Looking at the frequencies provided some initial insights. Most of the 1,555 people favored using the budget surplus for social security. This fact alone is not very interesting until you notice that a significant majority also favored using the surplus for tax cuts. This could be interpreted to mean that most Americans believe in having their cake and eating it too.

The survey included slightly more women that men. The largest category by race was white, followed by black and Hispanic at about 10% and 4% of the respondents. Of the social classes, there was an even split between people who described themselves as middle class and working class, with just a few ascribing themselves to the upper middle class. No one considered him or herself to be a member of the upper class.

It may also be surprising that our derived variable RACE_BIAS was positive for 75% of the respondents. This means that 3 out of 4 people indicated that they perceive that members of different racial groups are more or less trustworthy, intelligent and hard-working than other groups.

Seventy-five percent of the interviewees voted in the 2000 election. Although this was higher participation than the national average, it seems odd that after spending hours telling a survey taker how they think the country should be run, 25% of people chose not to express that information through the political process.

RULES

After reviewing the straight frequencies, clicking back to the Rules tab provides information about the relationships between the items. By default, the rules are expressed as "item A implies item B", and are listed with the following measures:

Expected confidence is the percentage of times item B occurs in the data.

Confidence is the percentage of cases in which item B is present when item A is present.

Support is the percentage of records containing both item A and item B.

Lift is how much more likely item B is if item A happens. A rule has lift when its confidence is higher than its expected confidence.

Count is the frequency of item A and item B occurring together.

The Enterprise Miner documentation notes that "A creditable rule has a large confidence factor, a large level of support, and a value of lift greater than 1" (SAS Institute, 2001). Berry and Linoff note that "lift comes closest to being useful on its own because it measures the extent to which the rule improves our ability to predict the right-hand side" (Berry, 2000).

The table in Figure 3 shows a small selection of the rules that were generated from the analysis of demographics, party affiliation and information sources.

EXP_ CONF	CONF	SUP- PORT	LIFT	COUNT	RULE
11.70	24.60	8.87	2.10	138	TALK RADIO ==> WATCHED SPEECHES & REPUBLICAN
24.76	38.28	5.14	1.55	80	65+ ==> EARLY LOCAL NEWS & DEMOCRAT
56.66	72.15	20.32	1.27	316	DAY TALK SHOW ==> FEMALE
62.57	75.11	11.45	1.20	178	18 - 29 ==> INTERNET
62.57	72.98	15.11	1.17	235	30 - 39 ==> INTERNET
62.57	72.67	7.01	1.16	109	WEST & INDEPENDENT ==> INTERNET
74.66	84.69	11.38	1.13	177	65+ ==> NEWSPAPER
74.66	83.69	7.59	1.12	118	MIDWEST & DEMOCRAT ==> NEWSPAPER
73.76	80.00	5.92	1.08	92	BLACK & DEMOCRAT ==> NATL NEWS

Figure 3.

The Association Node can produce huge numbers of rules with no guarantee that any of them will be valuable. And there is no automated technique that can separate the useful results from the common sense and inconsequential. The rules must be scanned and evaluated by someone who knows a new and actionable rule when they see it.

The rules involving many items and strong relationships tend to be obvious and not interesting. Most people would not be surprised by the third rule that daytime talk show viewers tend to be female, and only people who have never heard of Rush Limbaugh would be surprised by the first rule. We must drop down to rules with lower support to find clues to more interesting relationships, such as Independent voters in the West are more likely to get political information from the Internet, or that Midwestern democrats tend to read the newspaper.

The number and types of rules generated can be controlled by setting limits on the General tab of the Association Node, but this can have drawbacks too. We increased the minimum level of support for rules from the default of 5%. Although this reduced the number of rules to a more manageable list, it also eliminated all rules with an antecedent of BLACK. We decided to restore the lower limit because, even though those rules had low support, they had relatively high confidence and lift. And even though it may not be surprising that black respondents were very likely to be democrats, this fact might suggest ideas to explore further. Do democrats who are black favor the same policy issues as democrats who are white or Hispanic?

We initially limited the analysis to three items in an association, but when we expanded the limit to four items in the analysis of demographics vs. policy issues, we found one of the strongest rules, having a lift of 2.26, support of 8.6% and confidence of 50%. This four-item rule predicts that women aged 30-44 support stricter gun controls and have children at home.

PORTRAITS OF THE PARTIES

Using the View - Subset Table tool on the Enterprise Miner toolbar enabled us to create a quick portrait of people who described themselves as Democrat, Republican and Independent.

The highest-lift rules for Republicans tended to repeat the same items: tax cuts, school vouchers, death penalty, college-educated, middle to upper-middle class, as shown in Figure 4.

CONF	SUP- PORT	LIFT	_RHAND
40.30	10.42	1.87	WHITE & USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS
32.34	8.36	1.87	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & COLLEGE
22.14	5.72	1.83	USE SURPLUS FOR TAX CUTS & MIDDLE CLASS & FOR SCHOOL VOUCHERS
15.92	4.12	1.79	SOUTH & FOR SCHOOL VOUCHERS & COLLEGE
15.67	4.05	1.79	WHITE & UPPER MID CLASS & FOR DEATH PENALTY
16.92	4.37	1.78	WHITE & RURAL & FOR SCHOOL VOUCHERS
22.39	5.79	1.78	USE SURPLUS FOR TAX CUTS & MALE & FOR SCHOOL VOUCHERS
36.32	9.39	1.78	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & FOR DEATH PENALTY
16.42	4.24	1.77	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & AGE 45 ╤ 64
18.91	4.89	1.76	WHITE & SOUTH & FOR SCHOOL VOUCHERS

Figure 4.

Top 10 rules for democrats, by lift, involved black, age 45-64, and urban, and support of gun control, abortion rights, environmental protection and gays in the military, as listed in Figure 5

.

CONF	SUP- PORT	LIFT	_RHAND
19.96	6.75	2.13	USE SURPLUS FOR SOC. SEC. & BLACK
21.86	7.40	2.10	BLACK
15.40	5.21	2.08	USE SURPLUS FOR TAX CUTS & BLACK
15.02	5.08	2.05	RACE BIASED & BLACK

16.92	5.72	1.74	FOR GUN CONTROL & FOR ABORTION RIGHTS & AGE 45-64
15.59	5.27	1.65	URBAN & FOR GUN CONTROL & FOR ABORTION RIGHTS
17.49	5.92	1.62	URBAN & FOR GUN CONTROL & FOR ENVIR. PROTECTION
19.20	6.50	1.58	URBAN & NO CHILDREN AT HOME & FOR GUN CONTROL
17.87	6.05	1.58	USE SURPLUS FOR SOC. SEC. & FOR ABORTION RIGHTS & AGE 45 - 64
21.10	7.14	1.58	NO CHILDREN AT HOME & HS OR LESS & FOR GUN CONTROL

Figure 5.

Rules for Independents are listed in Figure 6. The rules with higher lift showed Independents tend to be younger and less educated and tend not to vote. They favored a mixture of issues including death penalty, gays in the military, and tax cuts.

CONF	SUP-PORT	LIFT	_RHAND
16.91	6.82	1.48	USE SURPLUS FOR TAX CUTS & NONVOTER & FOR GAYS IN MILITARY
18.02	7.27	1.47	USE SURPLUS FOR TAX CUTS & NONVOTER & FOR DEATH PENALTY
15.15	6.11	1.47	WHITE & NONVOTER & HS OR LESS
16.43	6.62	1.47	FOR DEATH PENALTY & AGE 18 - 29
16.11	6.50	1.46	RACE BIASED & NONVOTER & HS OR LESS
19.14	7.72	1.45	USE SURPLUS FOR TAX CUTS & RACE BIASED & NONVOTER
18.02	7.27	1.44	WHITE & USE SURPLUS FOR TAX CUTS & NONVOTER
20.41	8.23	1.44	RACE BIASED & NONVOTER & FOR DEATH PENALTY
16.75	6.75	1.44	RACE BIASED & AGE 18 - 29

Figure 6.

While some of the Democrat rules have high lift, they do not have very high confidence compared to the list of top Republican rules. This pattern is illustrated by graphical representations of the rules related to the three parties, shown in Figures 7, 8 and 9. The higher confidence for Republican rules suggests that Republicans may share similar backgrounds and a consensus of views within their party more than Democrats or Independents do.

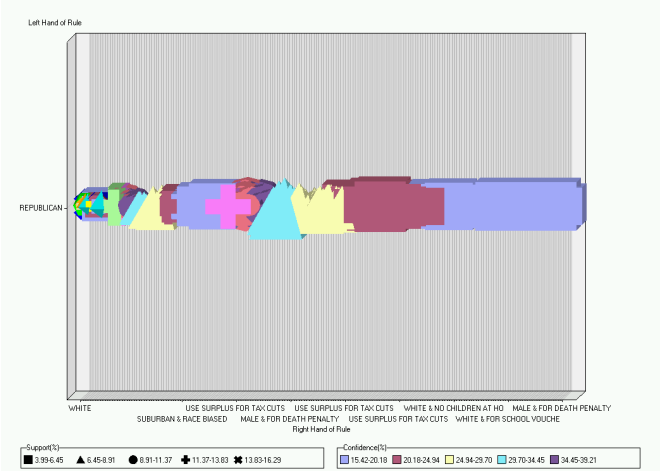

Figure 7. Plot of Republican Rules

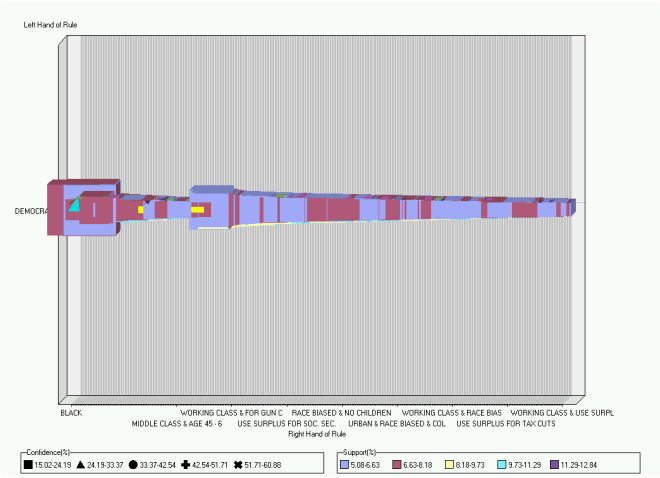

Figure 8. Plot of Democrat Rules

Independent

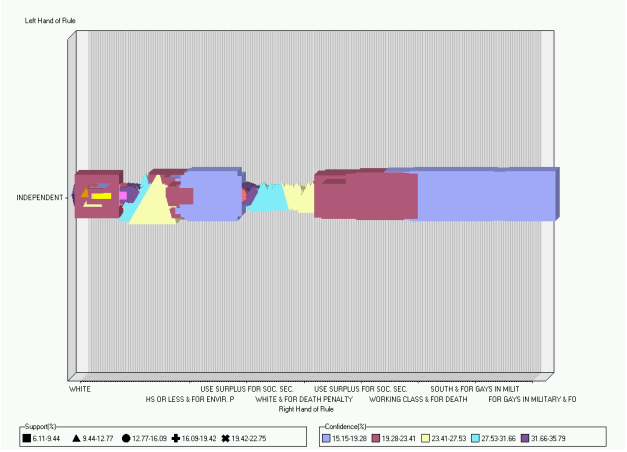

Figure 9. Plot of Independent Rules

CONCLUSION

Generating association rules can be a useful starting point for exploring unfamiliar data. Rules can help uncover interesting patterns that merit further examination. The results of association analysis will be more useful if the data is prepared so that the rules are expressed in plain English and if values with small frequencies are aggregated into meaningful groupings.

REFERENCES

Benisek, R., Giano, S., MacDougall, M., Mulvihill, P. and Ross, D. (2002) Data Mining 2000 Election Results (unpublished)

Berry, M. J. A. and Linoff, G. (2000) Mastering Data Mining: The Art and Science of Customer Relationship Management, New York: John Wiley & Sons, Inc.

Berry, M. J. A. and Linoff, G. (1997) Data Mining Techniques for Marketing, Sales and Customer Support, New York: John Wiley & Sons, Inc.

Burns, N., Kinder, D. R., Rosenstone, S. J., Sapiro, V. and the National Election Studies. American National Election Study, 2000: Pre- and Post-Election Survey [Computer file]. 2nd ICPSR version. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer], 2001. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002. http://www.icpsr.umich.edu

SAS Institute, Inc. (2001) Getting Started with Enterprise Miner [computer file], Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The author would like to acknowledge Ray Benisek, Sheila Giano, Paul Mulvihill and Dan Ross for their work in mining the 2000 Election Study, and Greg James for offering Cleveland-area students an opportunity to learn data mining techniques in a challenging classroom environment.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mary MacDougall
SPS Software Services, Inc.
2223 Chestnut Rd.
Cleveland, OH 44131
Email: marymacd@spssoft.com
Web: www.spssoft.com