Paper SA02_05

Data Mining Methods to Examine Thousands of Possibilities in Categorical Data

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

Categorical data are difficult to analyze when there are hundreds and thousands of possible values. Traditional statistical methods are inadequate. One way to reduce the number of values is to classify them into larger, broader categories using domain knowledge. Automatic reduction can be performed using SAS Text Miner. It is the purpose of this paper to provide several examples of category reduction using Text Miner, and to show how this reduction can aid in the analysis of data.

There are several possible ways to examine relationships in the categories when they are related to customer purchases. The first is to use PROC TRANSPOSE and CONCAT to link all purchases into one text string for each individual customer. This can be done including a Time ID as well. After the text strings are created, they can be analyzed using Text Miner. A second method is to decide beforehand on the maximum number of categories that can be handled easily in the data, and to use Text Miner to make the reduction so that other analyses such as association can be performed. Results to date have been very promising when applied to healthcare data.

INTRODUCTION

Given a set of transactions where each transaction is a set of items, an association rule is an implication of the form $X \rightarrow Y$ where X is the set of antecedent items and Y is the consequent item. In addition to the antecedent X and the consequent Y, an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis, the antecedent and consequent are sets of items called item sets that are disjoint ($X \cap Y = \emptyset$). The first number is called the support for the rule. It is the number of times that the combination appears. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent.

Another name for association rules is market basket analysis. However, market baskets are often representative of thousands or millions of items. There are too many different possibilities to work well with association rules. In many cases, it is possible to reduce the number of categories by defining broader classes containing many categories. For example, consider cereal, bread, cat food as representing many different items and brands. However, there are now so many different such classes defined that there still remain thousands of categories to examine and a market basket analysis is still difficult. As an example, on amazon.com, the DVD's that are associated with *The Passion of the Christ* are

- The Day After Tomorrow (Widescreen Edition) DVD
- DVDs from *The Spider-Man Feature Film Series*.
- DVDs from *The Harry Potter Feature Films Series*.
- DVDs from *The Kill Bill Series*.

even though the audiences for these items are very different. Nor has this association changed in a year's time. Drill down found the following additional DVD's: *Shrek II, The Day after Tomorrow, Star Wars Trilogy, Cold Mountain, Fahrenheit 9/11, Troy*.

Those who purchased *The Lord of the Rings* also purchased

- The Ultimate Matrix Collection (The Matrix / Reloaded / Revolutions / Revisited / The Animatrix) DVDs from *The Lord of the Rings (Extended Edition) Series*.
- DVDs from The Harry Potter Feature Films Series.
- DVDs from The Spider-Man Feature Film Series.

A pattern seems to be emerging. For *I, Robot*:

- The Day After Tomorrow (Widescreen Edition) DVD
- DVDs from *The Harry Potter Feature Films Series*.
- DVDs from *The Shrek Series*.

The general category is clearly science fiction and fantasy. However, that does not explain how *Cold Mountain, Kill Bill*, and *Fahrenheit 9/11* entered into association. It seems clear that broad classes of movies were defined, each containing many different individual films, and associations were developedacross the classes. Many of these

classes overlap. A similar search on “John Wayne” indicates that the actor forms a class by himself. His films are also in the class, “westerns” even if they are not westerns.

THE DATASET UNDER INVESTIGATION

The data used for this particular problem are in the public domain, downloaded from http://www.meps.ahrq.gov/Puf/PufDetail.asp?ID=149. The particular dataset used was MEPS HC-059A: 2001 Prescribed Medicines File. The dataset provides information on household-reported prescribed medicines for a nationally representative sample of the non-institutionalized population of the United States. The purpose of the dataset is to make estimates of prescribed medicine utilization and expenditures. Each record represents one prescribed medicine that was obtained during the year 2001. There are codes for individual households, and for individuals within each household. The dataset contains approximately 277,000 records. A partial list of the medications is given in Figure 1.

Figure 1. Partial List of Medications in the Dataset Figure

MEDICATION NAME (IMPUTED)

RXNAME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-7	201	0.07	201	0.07
-9	3246	1.17	3447	1.24
7 CYCLES NORDETTE	1	0.00	3448	1.24
A & D OINTMENT	11	0.00	3459	1.24
A/B OTIC	36	0.01	3495	1.26
ABSORBASE	8	0.00	3503	1.26
ACCOLATE	146	0.05	3649	1.31
ACCU-CHEK	3	0.00	3652	1.31
ACCU-CHEK ADVANTAGE	5	0.00	3657	1.32
ACCU-CHEK ADVANTAGE (STRIP)	8	0.00	3665	1.32
ACCU-CHEK ADVANTAGE (STRIP,2X50)	5	0.00	3670	1.32
ACCU-CHEK ADVANTAGE CARE	6	0.00	3676	1.32
ACCU-CHEK COMFORT CURVE (STRIP)	5	0.00	3681	1.32
ACCU-CHEK COMFORT CURVE STRIPS	7	0.00	3688	1.33
ACCU-CHEK INSTANT	1	0.00	3689	1.33
ACCU-CHEK SIMPLICITY CARE (COMPLETE MONITORING)	4	0.00	3693	1.33
ACCU-CHEK TEST STRIP	7	0.00	3700	1.33
ACCUPRIL	1507	0.54	5207	1.87
ACCURETIC (10X3)	29	0.01	5236	1.88
ACCUTANE	72	0.03	5308	1.91
ACCUTANE (RX PAK, 10X10)	42	0.02	5350	1.93
ACCUZYME	1	0.00	5351	1.93
ACEBUTOLOL HCL	89	0.03	5440	1.96
ACEON	42	0.02	5482	1.97
ACEPHEN	6	0.00	5488	1.98
ACETAMIN	19	0.01	5507	1.98
ACETAMIN W/COD	1	0.00	5508	1.98
ACETAMIN/COD3	28	0.01	5536	1.99
ACETAMINAPHEN	1	0.00	5537	1.99
ACETAMINOPHEN	358	0.13	5895	2.12
ACETAMINOPHEN (A.F.,CHERRY)	8	0.00	5903	2.12
ACETAMINOPHEN (DROPS)	4	0.00	5907	2.13
ACETAMINOPHEN (DROPS, A.F.)	1	0.00	5908	2.13
ACETAMINOPHEN (INFANT)	1	0.00	5909	2.13
ACETAMINOPHEN E.S.	1	0.00	5910	2.13
ACETAMINOPHEN INFANTS	1	0.00	5911	2.13
ACETAMINOPHEN INFANTS (DROPS,A.F.,FRUIT)	1	0.00	5912	2.13
ACETAMINOPHEN W/ CODEINE	10	0.00	5922	2.13
ACETAMINOPHEN W/COD	36	0.01	5958	2.14
ACETAMINOPHEN W/CODEINE	25	0.01	5983	2.15
ACETAMINOPHEN W/CODEINE #3	24	0.01	6007	2.16
ACETAMINOPHEN W/CODEINE ELIXIR	1	0.00	6008	2.16
ACETAMINOPHEN W/CODEINE#3	1	0.00	6009	2.16
ACETAMINOPHEN WITH CODEINE	1	0.00	6010	2.16
ACETAMINOPHEN/APAP	1	0.00	6011	2.16

A quick review of the different medication names demonstrates very clearly that there are many medications that are virtually identical, with only slight changes. Although names are somewhat standardized, there is enough discretion in data input to have some names occur only occasionally as they are just slightly different from the norm. Misspellings or abbreviations are also apparent, “acetamin”. Numeric codes are used for unknown or missing medications (-7,-9). It would be very helpful to have an easy way to simplify the list. For the most part, the difference in drug name occurs because of the level of detail provided at data entry.

NEED FOR COMPRESSING THE DATASET

To run the association node, a target variable and an ID variable need to be in the dataset. In EM 5.1, the dataset also needs to be identified as transactional. In this particular dataset, the household is an identifier; a second identifier is that of individual within the household. One of the problems with having too many categories in using the Association Node in Enterprise Miner is the following error:

```
61488       options nocleanup;
```

```
61489       Proc Assoc dmdbcat= EMPROJ.dm_DGM00006
61490                  data= _emtrain
61491                  out=EMDATA.ASCGV5V6 (label = "Output from Proc Assoc")
61492                    items=4;
61493            customer
61494                MD
61495            ;
61496            target
61497                DIAGNOSIS
61498         ;
61499              run;
----- Potential 1 item sets = 5405 -----
Counting items, records read:    14242
Number of customers:                24
Support level for item sets:         1
Maximum count for a set:            14
Sets meeting support level:       5405
Megs of memory used:              4.10

----- Potential 2 item sets = 14604310 -----
Counting items, records read:    14242
Maximum count for a set:            12
Sets meeting support level:    2175908
Megs of memory used:            588.28
Error: Out of memory.  Memory used=1642.0 meg.

Item Set 3 is null.
NOTE: The SAS System stopped processing this step because of insufficient memory.
WARNING: The data set EMDATA.ASCGV5V6 may be incomplete.  When this step was stopped there
were
         2181314 observations and 6 variables.
WARNING: Data set EMDATA.ASCGV5V6 was not replaced because this step was stopped.
NOTE: PROCEDURE ASSOC used (Total process time):
    real time       49.98 seconds
```

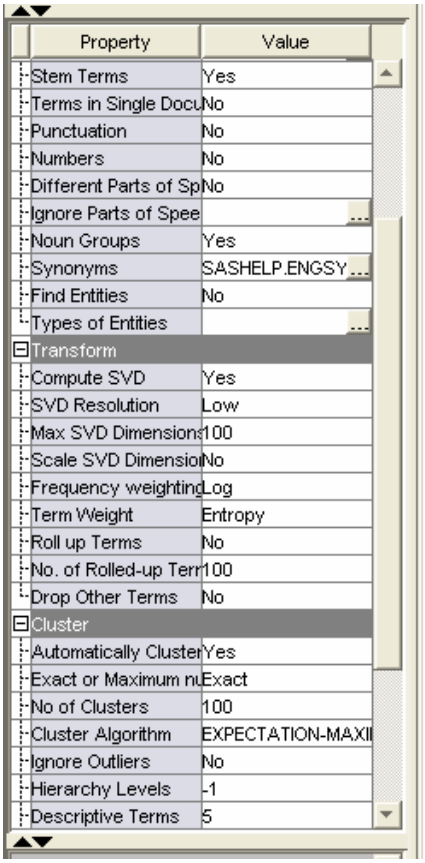
Too many categories results in too many rules, most of them meaningless. While it is possible to change the defaults, it is preferable to reduce the number of categories. The instructions for changing the defaults is provided in the SAS Help Documentation:

> When the Association node processes a data set that contains variables with a very large number of classes, it compares the number of levels to two values: the Association node class variable threshold value of 100,000, and the value stored in the macro variable DM_MAX_TRAIN_LEVELS. **If the number of levels in a class variable exceeds the greater of the two values, Enterprise Miner generates an error and halts the process flow.**
>
> To use the DM_MAX_TRAIN_LEVELS macro variable to control variable levels in transaction data with the Association node, it must be set to some value greater than 100,000. (The default setting is 512.) To set the macro variable to a new value of 120,000, for example, type the following in the SAS Program Editor: %let DM_MAX_TRAIN_LEVELS = 120000 and then submit the statement to SAS for processing.

RESULTS OF ASSOCIATION RULES ON MODIFIED DATASET

There is a way to compress these categories using the similarity in the names of the medications by using Text Miner. Since Text Miner clusters text information, the medication names can be used to define clusters. In the text mining process, words with similar stems are considered similar and will be placed in the same cluster. The investigator can decide beforehand on an upper limit to the number of clusters. The methodology in EM 5.1. is given in Figure 2.

Figure 2. Icons for Text Mining to Reduce Clusters



The defaults for Text Miner are modified slightly. In the Cluster section, the default for "Automatically Cluster" is changed to Yes. The value for "Exact or Maximum" is changed to "Exact" and the number of clusters is specified to whatever the investigator wants them to be. Once the clustering is complete, the datasets defined in Text Miner are merged to retain the descriptors of the cluster as well as the relationship of the cluster to the identifier. Once the merge is complete, the dataset can then be examined using the Association Node.

```
data sasuser.clusternamescopy (keep=_cluster_ _freq_ rmsstd_ clus_desc);
set emws.text_cluster;
run;
data sasuser.desccopy (drop= _svd_1 _svd_100 prob1-prob100 );
set emws.text_documents;
run;
proc sort data=sasuser.clusternamescopy;
by _cluster_;
proc sort data=sasuser.desccopy;
by _cluster_;
data sasuser.medicationswithdescriptions;
merge sasuser.clusternamescopy sasuser.desccopy;
by _cluster_;
run;
```

A clustering limited to 100 categories is given in Table 1.

Table 1. Representative Clusters

#	Descriptive Terms	Freq	Percentage
1	celexa, flovent, k-dur, lorazepam, hctz/triamterene	83273	30%
2	w/applicator, + estrogen, vaginal, estrace, clotrimazole	6438	2%
3	diflucan, serevent, ventolin, diskus, advair	2880	1%
4	nestabs, accutane, nephro-vite, premesis, cbf	327	0%
5	ventolin, inhaler, intal, refill, nebulizer	3848	1%
6	diclofenac, sodium, acetaminophen/codeine, warfarin, dicloxacillin	4967	2%
7	claritin, reditabs, redi, + unit	2890	1%
8	fruit, gnp, pediatric, gantrisin, pediatric fruit	529	0%
9	zoloft	2407	1%

#	Descriptive Terms	Freq	Percentage
10	bitartrate, apap/hydrocodone, guiatuss, baby, miralax	1131	0%
11	pink, petrolatum, napsylate, apap/propoxyphene, propoxyphene	738	0%
12	fluticasone, propionate, clobetasol, prop, halobetasol	111	0%
13	srn, ml, insulin, nph, human	2174	1%
14	zyrtec, d.f.,s.f.,banana-grape, cetirizine	1682	1%
15	coumadin, kapseals, infatabs, kapseal, dilantin	2052	1%
16	cozaar, evista, hyzaar, lasix, claritin-d	7828	3%
17	caplet, pd, bromfenex, gen, calan	194	0%
18	df, strawberry, dihistine, af, strawb/pineap	90	0%
19	dialpak, ortho, tri-cyclen, ortho-novum, ortho-cyclen	1948	1%
20	+ packet, single, emla, zpak, cholestyramine	1379	0%
21	wellbutrin, cardizem, provera, maxzide, elavil	2535	1%
22	atenolol, tenormin	4549	2%
23	+ strip, ultra, natalcare, precision, test	109	0%
24	celebrex	3108	1%
25	nystatin, cherry	4	0%
26	strawberry, cefaclor, light, locholest, tannihist	121	0%
27	furosemide	3321	1%
28	hydrochlorothiazide	3285	1%
29	cromolyn, heparin, dosette, cefazolin, vial	111	0%
30	prempro, + dispenser, alesse, ez-dial, minipack	2557	1%
31	maxalt, caplet, + unit	78	0%
32	verapamil, amitriptyline, trazodone, clonidine, cyclobenzaprine	15295	6%
33	synthroid	4243	2%
34	toprol	1723	1%
35	zestril, lisinopril	2956	1%
36	metformin, xr, diltiazem	126	0%
37	+ multivitamin, compound, bactrim, anaprox, aciphex	2206	1%
38	zocor, + unit	1347	0%
39	pilocarpine, drop-tainer	9	0%
40	ins, bd, terumo, + syringe, sodium chloride	35	0%
41	vial, p.f., albuterol, + sulfate, s.d.	722	0%
42	one, touch, + lancet, syr, point	2994	1%
43	metoprolol, tartrate, succinate, loxapine, brimonidine	1877	1%
44	ex, gelcaplet, fa, dilantin, chromagen	24	0%
45	remeron, sinemet, dynacirc, eskalith, norpace	565	0%
46	filmtab, fruit, nr, hc, s.f.	466	0%
47	patanol, hydrochloride, buspar, phoslo, codeine/promethazine	1637	1%
48	zyban, vivelle-dot, daily, vivelle, advantage	641	0%
49	prevacid	2291	1%
50	allegra	1850	1%
51	metformin, xr, effexor, adderall, glucophage	4419	2%
52	glyburide, retin-a, micro, retin-a micro, copley-d	1903	1%
53	ibuprofen, ibu, grx, motrin, berry	2496	1%
54	levoxyl	1705	1%
55	+ unit, prilosec, omeprazole, nf-omeprazole	2923	1%

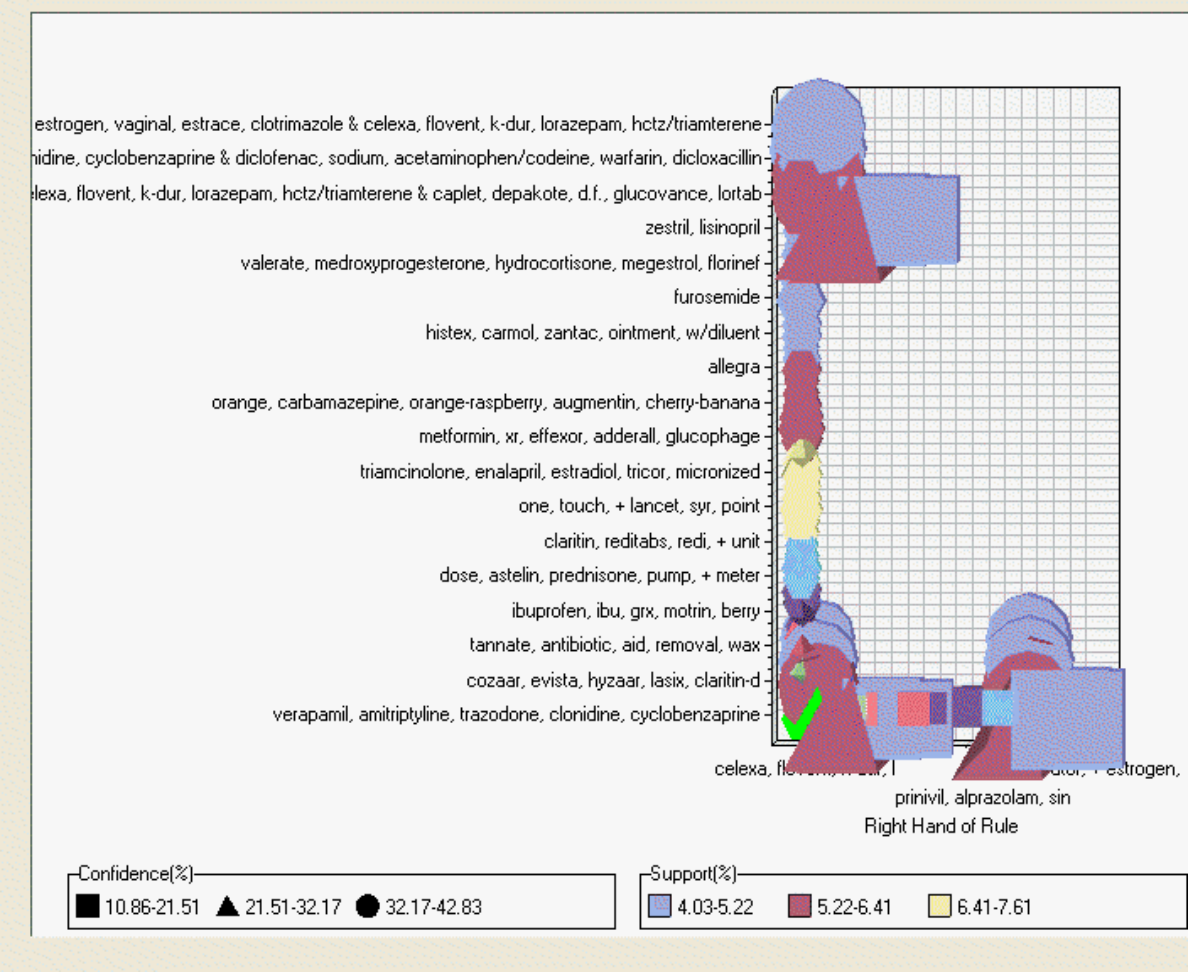
#	Descriptive Terms	Freq	Percentage
56	sulfa, silver	6	0%
57	prevident, roxicet, antacid, gel-kam, therapy	2010	1%
58	paroxetine	33	0%
59	tannate, antibiotic, aid, removal, wax	2231	1%
60	vioxx, tazorac, rofecoxib, estring, nf	2759	1%
61	zocor, package, bulk, simvastatin	1948	1%
62	gum, amoxil, strawberry, amoxicillin, trihydrate	2906	1%
63	clonazepam, package, bulk, mevacor, pepcid	762	0%
64	methylprednisolone	75	0%
65	orange, carbamazepine, orange-raspberry, augmentin, cherry-banana	1290	0%
66	la, nifedipine, detrol, theophylline, adalat	3303	1%
67	histex, carmol, zantac, ointment, w/diluent	1240	0%
68	bitartrate, apap/hydrocodone, caplet	156	0%
69	valerate, medroxyprogesterone, hydrocortisone, megestrol, florinef	1455	1%
70	junior, orange, motrin	2	0%
71	prinivil, alprazolam, singulair, package, bulk	5862	2%
72	fresh, + pad, skin, ery, prep	67	0%
73	lithium, carbonate, haloperidol, oyster, atorvastatin	788	0%
74	ranitidine, n/apap, propoxy, propoxyphene-n, propoxyphene-n w/apap	69	0%
75	blister, micardis, blister card,4x7, acid, blister pack,6x28	1206	0%
76	erythromycin, mesylate, doxazosin, citrate, mononitrate	3634	1%
77	vial, humalog, pen, ipratropium, bromide	2738	1%
78	allergy, allergy relief medicine, medicine, allergy relief, relief	24	0%
79	accupril	1507	1%
80	digoxin, lanoxin, valium, diazepam, glipizide	26	0%
81	saline, + drop, broncho, q-pap, floxin	1309	0%
82	orange, chloride, potassium chloride, penicillin, oxybutynin	2283	1%
83	+ sulfate, morphine, ferrous sulfate, quinine, ferrous	2691	1%
84	tamiflu, amerge, compack, estrostep, demulen	1707	1%
85	imitrex, nitroglycerin, glucometer, elite, care	1426	1%
86	paxil, + unit, paroxetine, orange	2361	1%
87	pravachol	1489	1%
88	caplet, depakote, d.f., glucovance, lortab	5963	2%
89	b-d, + syringe, cannula, bd, ins	27	0%
90	lipitor	5970	2%
91	formula, prenatal vitamins, wildberry, natachew, prenatal formula	373	0%
92	cough, control, mytussin, s.f.,fruit/mint, benzoyl	327	0%
93	clindamycin, phosphate, pledget, phos, clindets	154	0%
94	blister, amerge, + unit, pck, pk	1480	1%
95	non-aspirin, + child, gra, diphenhist, a.f.,fruit	108	0%
96	insulin, srn, humulin, novolin, mcg	269	0%
97	norvasc, amlodipine, besylate	3839	1%
98	dose, astelin, prednisone, pump, + meter	2983	1%
99	triamcinolone, enalapril, estradiol, tricor, micronized	3239	1%
100	guaifenesin, w/codeine, w/codeine #3, phenergan, vc	2654	1%

An individual with domain knowledge would have to validate the clusters by investigating their reasonableness. Consider cluster 55 where Prilosec is the brand name for omeprazole. As another example, Zestril (cluster 35) is the brand name for lisinopril. Similarly, Redi tab is a delivery form for Claritin (cluster 7). These examples give a strong indication that Text Miner can pick up alternate names for similar items even without defining a synonym list to equate them.

While at first glance cluster 3 appears to have a problem because Diflucan is a medication used to treat yeast infections while the remaining medications treat asthma, one of the side effects of the asthma medications is oral yeast infections. Cluster 1 is also puzzling since the different medications are used for very different diagnoses. It is one that must be investigated in more detail. This will be done using other aspects of Text Miner.

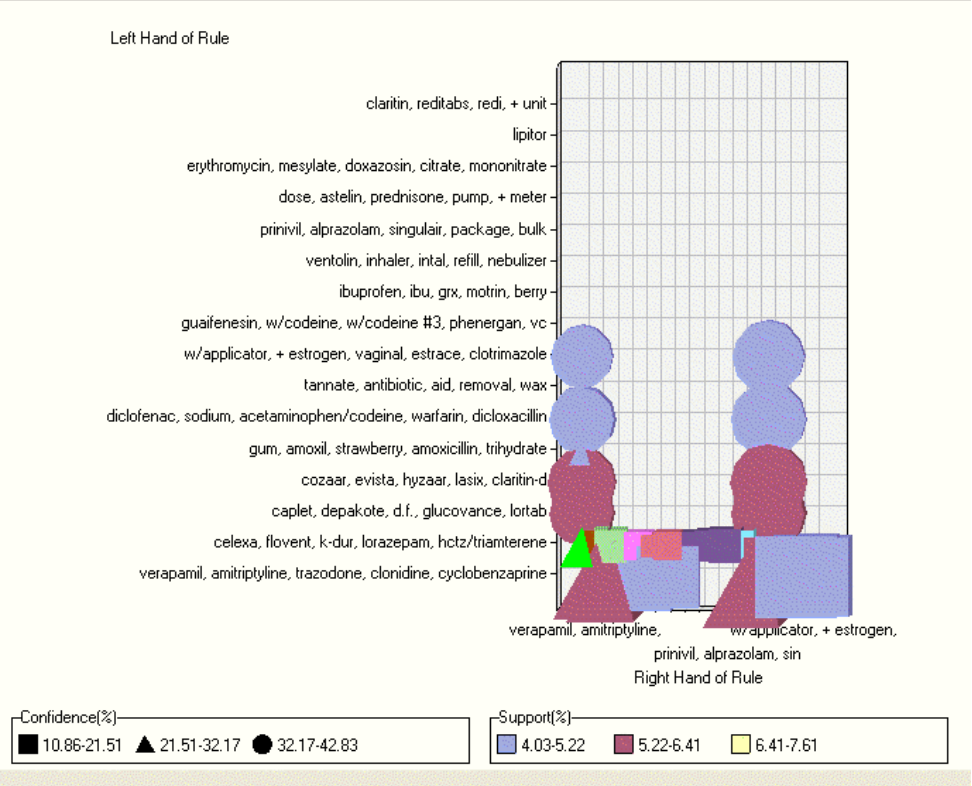
Once the categories are condensed, the Association Node is used, and a graph displayed of the results (Figure 3). There were a total of 84 rules defined with these 100 clusters.

Figure 3. Association Rules for 100 Clusters



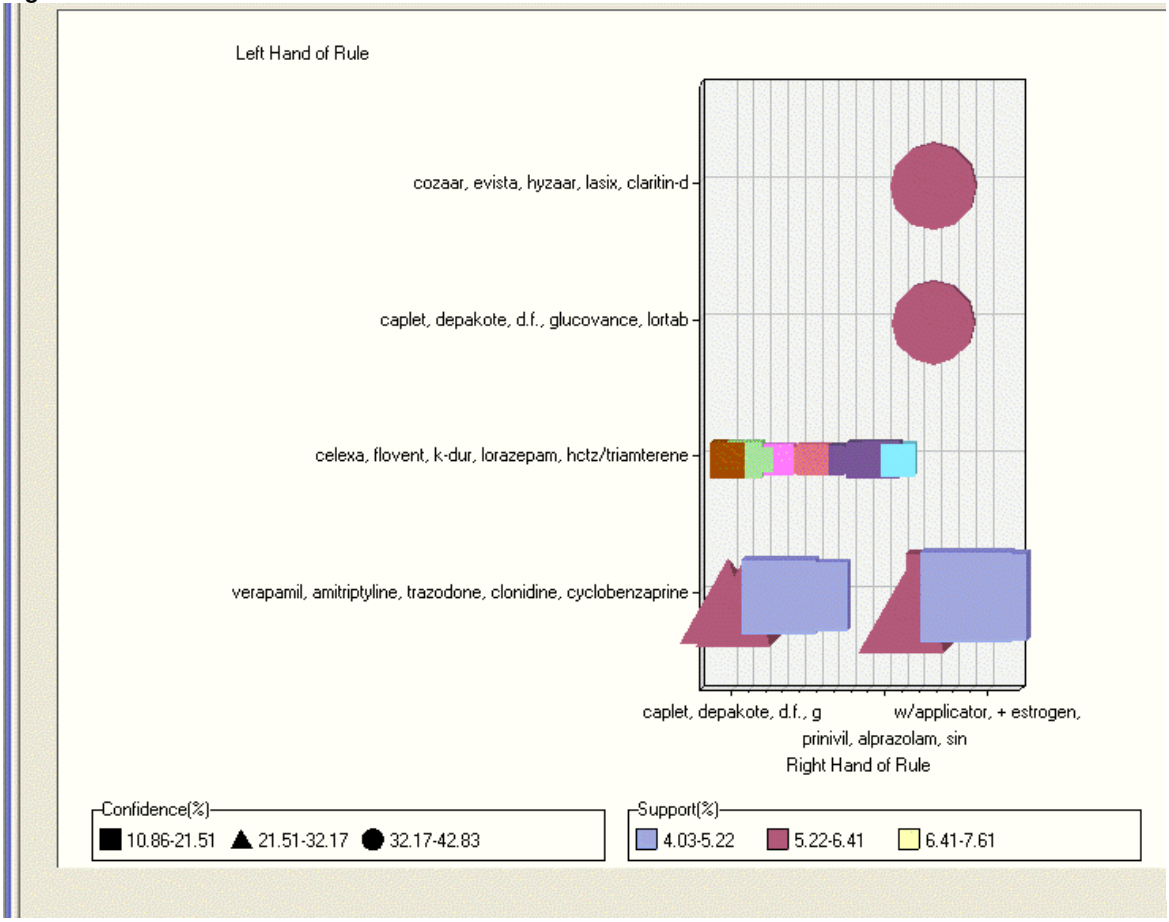
Note that there is one vertical line at the category, celexa (depression), etc,, on the right-hand side of the association rule. It appears that these are common medications used in combination with many other medications. Since several apparently disjointed medications are combined into one category that appears to be related to almost all other categories, it becomes important to investigate it in detail. Figure 4 takes advantage of the drill-down properties of the Association node to examine in more detail the lower portion of the graph in Figure 3. Figure 5 is an additional drill-down. This type of drill-down is only available in Version 4.3 of SAS Enterprise Miner; it is not available in Version 5.1.

Figure 4. Drill-Down into Association Rule Graphical Display



In this drill-down, there appears to be two values on the right-hand side that relate in the same way to the medications on the left-hand side.

Figure 5. Additional Drill-Down



Further drill-down allows the investigator to examine relationships more individually.

To examine the rules in more detail, the “Where” clause was used to find rules for the drug, Celexa (Figure 6).

Figure 6. Rules for Celexa

	Rule
1	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> verapamil, amitriptyline, trazodone, clonidine, cyclobenzaprine
2	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> caplet, depakote, d.f., glucovance, lortab
3	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> cozaar, evista, hyzaar, lasix, claritin-d
4	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> gum, amoxil, strawberry, amoxicillin, trihydrate
5	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> diclofenac, sodium, acetaminophen/codeine, warfarin, dicloxacillin
6	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> tannate, antibiotic, aid, removal, wax
7	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> w/applicator, + estrogen, vaginal, estrace, clotrimazole
8	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> guaifenesin, w/codeine, w/codeine #3, phenergan, vc
9	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> ibuprofen, ibu, grx, motrin, berry
10	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> ventolin, inhaler, intal, refill, nebulizer
11	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> prinivil, alprazolam, singulair, package, bulk
12	celexa, flovent, k-dur, lorazepam, hctz/triamterene ==> dose, astelin, prednisone, pump, + meter

Rule number 9 shows a relationship with pain medications, rule 11 with asthma medications. Rule 4 with flavorings listed indicates a relationship to pediatric patients. To contrast, there is only one rule for Zestril (Figure 7), and that is to Celexa on the right-hand side.

Figure 7. Association Rule for Zestril

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.11	4.19	88.12	445.00	zestril, lisinopril ==> celexa, flovent, k-dur, lorazepam, hctz/triamterene

USE OF SAS TEXT MINER FOR ASSOCIATION

There are a number of analyses available in Text Miner. The first is a list of terms in the variable field. In addition, association rules are visualized to examine the relationships between terms. In Text Miner, association rules define the strength of association between different terms in the wordlist. In order to use SAS Text Miner for Association, the form of the data must be changed. All medications for one household (or one patient, depending upon the interest) must be linked, and the observational unit must be changed from medication to household. The following code will make the required changes:

```
proc sort data = sasuser.icuchargesonly out= work.sort_out;
    by mr_ charges;
run;
data work.sort_out1;
    set work.sort_out;
            charges = translate(left(trim(charges)),'_',' ');
run;
proc Transpose data=work.sort_out1
            out=work.tran (drop=_name_ _label_)
               prefix=med_;
    var charges ;
    by mr_;
run;
data work.concat( keep= mr_ charges ) ;
            length charges $32767 ;
            set work.tran ;

            array chconcat {*} med_: ;

            charges = left( trim( med_1 )) ;

            do i = 2 to dim( chconcat ) ;
                        charges = left(trim(charges)) || ' ' || left(trim( chconcat[i] )) ;
            end ;
run ;
proc sql ;
            select max( length( charges )) into :charges_LEN from work.concat ;
quit ;
```

```
%put charges_LEN=&charges_LEN ;
data sasuser.icutextstrings ;
        length charges $ &charges_LEN ;
        set work.concat ;
run ;
proc contents data=sasuser.icutextstrings ; run ;
```

Once this code is run, an initial household record as shown in Table 2 is modified by the above code as shown in Table 3.

Table 2. Household Record

DUID	RXNAME
40001	SOFTCLIX
40001	ANTIVERT
40001	SOFTCLIX
40001	SOFTCLIX
40001	SOFTCLIX
40001	SOFTCLIX
40001	SOFTCLIX
40001	SOFTCLIX
40001	ESTROGEN
40001	ESTROGEN
40001	CEFADROXIL
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN
40001	ESTROGEN

Table 3. Modified Patient Record

```
DUID      RXNAME

40001     ANTIVERT CEFADROXIL ESTROGEN ESTROGEN ESTROGEN ESTROGEN ESTROGEN ESTROGEN
ESTROGEN ESTROGEN ESTRO BETAMETH KEFLEX
40007 ACCUTANE_(RX_PAK,_10X10) ACCUTANE_(RX_PAK,_10X10) ACCUTANE_(RX_PAK,_10X10) ALBUTEROL
BENZAMYCIN B
40010 ALBUTEROL APAP/PROPOXYPHENE_NAPSYLATE_(PINK) BENZTROPINE_MESYLATE CARDEC_DM_SYRUP
ERYTHROMYCIN/SU
```

Note that medications that use multiple words are linked by underscores to preserve the entire name in Text Miner. Text Miner is then used on the re-defined dataset to find the relationships on those text strings. The clusters are given in Table 4. below. By using a combination of medications that are linked, the optimal number of clusters as identified in Text Miner is equal to 21. Once the clusters are defined, they can be used with other statistical analyses. Association is preserved in the linkage defined by the SAS code. The Association Node can no longer be used with this dataset as each customer now has only one record. Attempts to run the Association Node will result in a finding of zero rules.

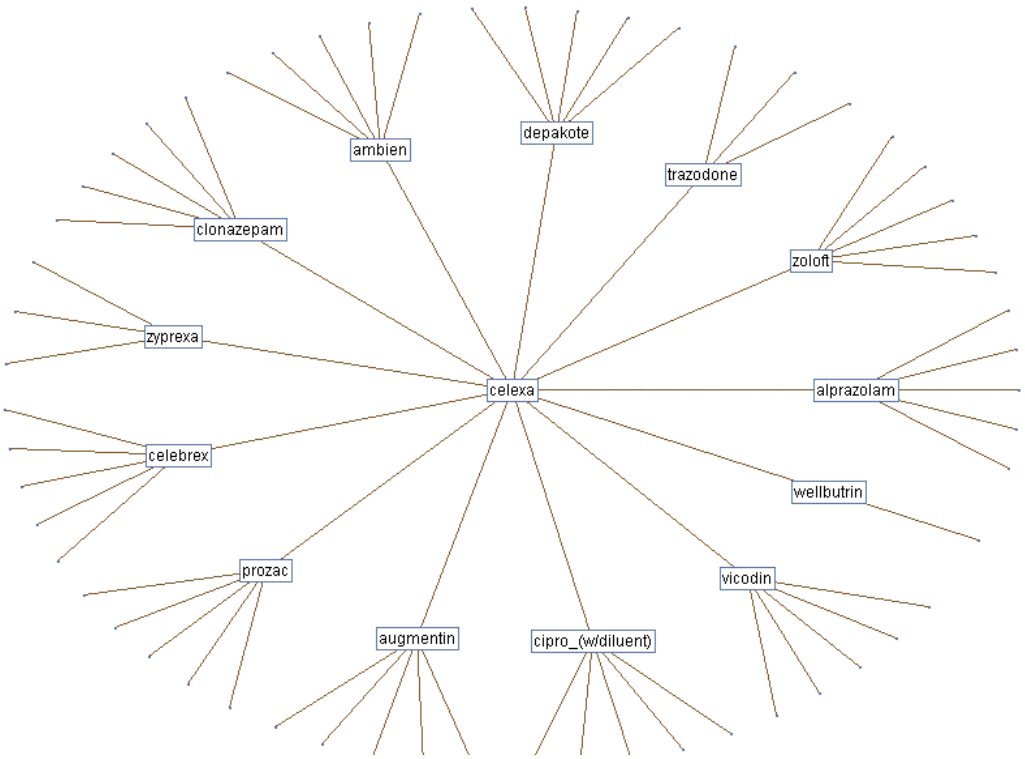
Table 4. Clusters of Medications

Cluster Number	Cluster Description	Cluster Label
1	glucophage, furosemide, zestril, synthroid, softclix	Diabetes
2	allegra, claritin, flonase, nasonex, hydrochlorothiazide	Allergies
3	lipitor, vioxx, allegra, claritin	Lipitor, arthritis and allergies
4	augmentin	Children's antibiotic
5	vicodin, apap/hydrocodone_bitartrate, celexa, naproxen, vioxx	Pain
6	zoloft, triple_antibiotic, paxil, zyrtec, naproxen	Depression, allergies

Cluster Number	Cluster Description	Cluster Label
7	synthroid, premarin, augmentin, zithromax, amoxicillin	Post-menopause and antibiotics
8	levoxyl, synthroid, premarin, lipitor, zoloft	Post-menopause, cholesterol, and Depression
9	hydrochlorothiazide, lipitor, norvasc, furosemide, zestril	Cholesterol, Hypertension
10	estradiol, prednisone, prempro, zocor, zyrtec	Post-menopause, allergies, cholesterol
11	glyburide , aspirin, glucophage, lisinopril, simvastatin	Diabetes, cholesterol
12	lanoxin, coumadin, pravachol, furosemide, norvasc	Heart, cholesterol
13	atenolol, hydrochlorothiazide, lipitor, zestril, premarin	Heart, cholesterol
14	fosamax_(unit_of_use, blister_pck, fosamax_(unit_of_use), fosamax, celebrex	Ulcer, arthritis
15	fluoxetine_hcl, prozac, zithromax, augmentin, cephalexin	Antibiotic, depression
16	singulair_(unit_of_use), albuterol, serevent, flovent, singulair	Asthma
17	celebrex, ultram, hydromet, vioxx, ibuprofen	Arthritis
18	premarin, medroxyprogesterone_acetate_(unit_of_use), allegra, triple_antibiotic, cephalexin	Post-menopause, antibiotics
19	prevacid, cephalexin, amoxicillin, ibuprofen, celebrex	Ulcer, antibiotic, arthritis
20	triple_antibiotic, augmentin, ibuprofen, motrin, amoxicillin	Antibiotic, pain
21	paxil, paxil_(unit_of_use), alprazolam, lorazepam, amoxicillin	Depression, antibiotic

In addition to clusters, concept links can be used to find relationships. Text Miner uses association rules to define the concept links. In order to be visualized, the relationship between terms is assumed by default to be highly significant. That is, the chi-square statistic is greater than 12. Both terms occur in at least n documents. The default value of n is MAX(4, A/100, B), where A is the largest value of NUMDOCS for the subset of terms that are used in concept linking, B is the 1000th largest value of NUMDOCS for the subset of terms that are used in concept linking, and NUMDOCS is the number of documents that a term appears in. For example, for 13,000 documents, every term in the term set must occur in at least 130 documents. Figure 5 shows the concept links for the medication, Celexa, that was originally combined with other medications as given in Table 1.

Figure 5. Concept Links for Celexa

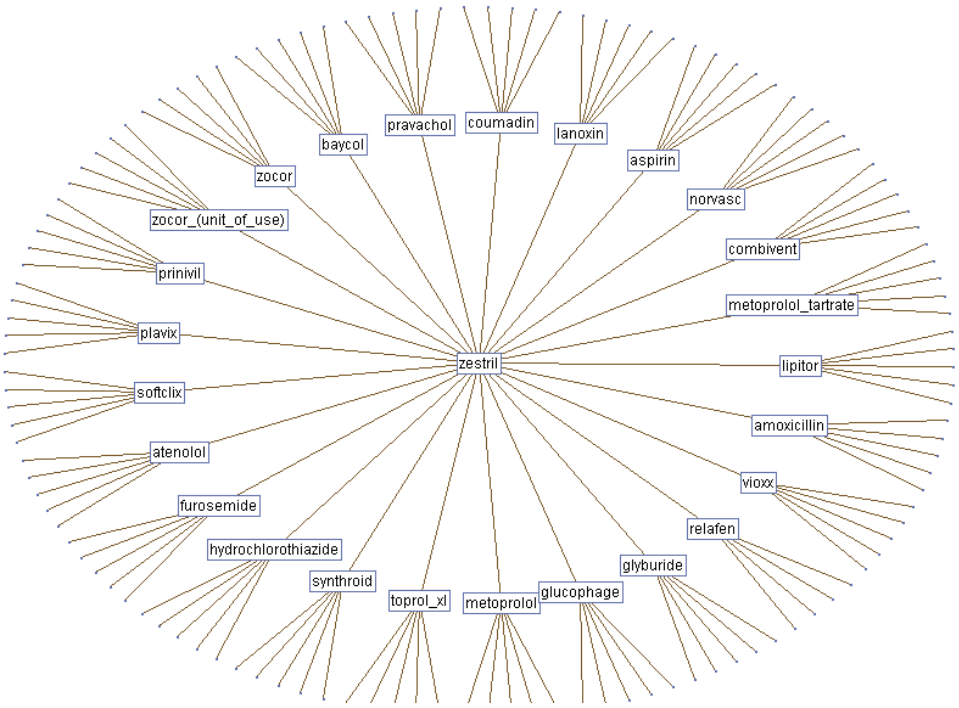


Note that the investigator controls the left-hand side of the rule while the right-hand side is allowed to vary. Since the concept link is created using Active-X, placing the cursor over one of the terms provides the number of documents containing both the term and the initial term Celexa divided by the number of documents containing that term.

In other words, the concept links provide the confidence value for the association. Note that the other medications listed in Cluster 1 in Table 1 such as Flovent are not included in this concept link. Instead, medications, such as Prozac and Zoloft are listed. These links indicate that patients do often switch from one drug to another either

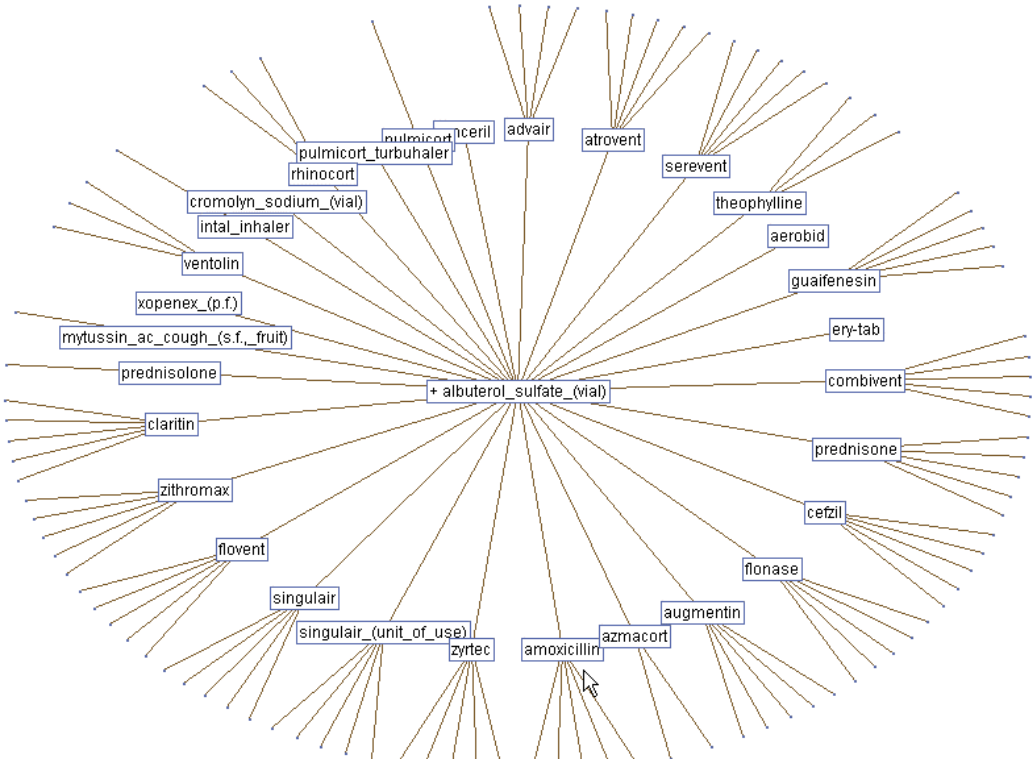
because the drug is not effective in treating the condition, or there is a change in insurance formularies. Also, note that the medication, Vicodin (a pain killer) is listed.

Figure 5. Concept Links for Zestril



In contrast to the results from the Association Node, there are many links to Zestril. They include many heart and diabetes medications, indicating a strong link. Zestril is often used to treat heart problems in combination with other medications.

Figure 6. Concept Links for Albuterol



Note that Albuterol is related to various other asthma medications such as Singulair, Zyrtec, and Serevent. It is also related to antibiotics such as Zithromax. There is a relationship between Albuterol and Flovent. However, in Table 1, Flovent is related to Celexa rather than to Albuterol. This is an indication that the relationship of Albuterol to Celexa is second generation.

CONCLUSION

Text Miner combined with the Association Node can be used to reduce a field containing hundreds and thousands of levels into a target variable that is much more reasonable to investigate. Once the number of clusters is reduced, they can be combined with other statistical methods to analyze the data.

Text Miner applies association rules through concept links. Linkage between items in a market basket is preserved through the construction of a text string where all purchases are contained within the string. The method is successful when the items in the market basket are “stemmed” so that similar items have similar base words.

ACKNOWLEDGMENTS

The author wants to acknowledge the help of the SAS Consultants, Darius S. Baer and Ross Bettinger for their help in coding the data compression.

CONTACT INFORMATION

Patricia B. Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6826
502-852-7132 (fax)
pcerrito@louisville.edu