A top-k analysis using multi-level association rule mining

for autism treatments

Introduction

Association rule mining, also called market-basket analysis, uses a number of generic measures to determine items that are associated with each other. This association does not imply causation although correlated items will be discovered during the mining process. The two most commonly used measures of interestingness to generate the rules are support and confidence. Support represents the percentage of all records where the items listed in the rule are present whereas confidence represents the accuracy of the rule.

Strong rules (i.e., rules that meet the minimum support/confidence thresholds) can be misleading since they may be negatively correlated. A separate correlation analysis is necessary in order to preserve only the rules that have a strong positive correlation. Other measures may be used to prune rules that are negatively correlated. Measures such as lift, leverage and conviction are used for these purposes. Chi-square tests can also be used to not only determine the direction of the correlation (positive or negative) but also the strength of that correlation.

Oftentimes in association rule mining it can be difficult to generate rules due to the sparsity of the data at the most atomic levels (Han and Kamber 2006). Multi-level association rule mining alleviates this problem by segmenting the data according to a user-defined hierarchy. In the IAN dataset, there are over 500 treatments although not all are commonly used. Therefore, it was difficult to find meaningful results at this atomic level. By segmenting the data records by treatment category, we were able to generate a top-K list of treatments.

Motivating Example

On average, an autistic child is on seven different treatments at any given point in time (Green, Pituch et al. 2006). The efficacy of the majority of these treatments is not supported by clinical trials with few exceptions including risperidone (Pandina, Bossie et al. 2006), Ritalin (Aman, Arnold et al. 2005), ABA therapy (Corsello 2005) etc. In terms of available pharmacological agents, only risperidone and Ritalin have replicable results for the autistic population (Parikh, Kolevzon et al. 2008).

There are more than 500 treatments currently in the dataset we are analyzing for treating autism; these treatments include medications, vitamin and mineral supplements, diets, behavioral/educational interventions as well as alternative treatments designed to treat the autistic child holistically. It would not be feasible or prudent to conduct clinical trials on such a large number of treatments. Data mining, for this particular scenario, would be helpful in reducing the solution space to determine the top 'k' treatments. The efficacy for this smaller group of treatments could then be properly confirmed in a clinical trial. Multiple analyses would be advantageous in pruning the solution space in order to overcome any limitations in one particular method (i.e., strong association rules that are negatively correlated).

The IAN web project is designed to serve as a national registry for autistic children and their families. Its purpose is two-fold: (1) to connect researchers to large, diverse subject bank and; (2) to collect data for secondary analysis and release the de-identified data to researchers with IRB approval. There are approximately 300,000 families in the United States that have at least one child between the ages of 3-18 that is afflicted with autism. The goal for the IAN web project is to register 100,000, or 1/3, of these families. This data was analyzed using a combination of association rules and statistical measures to find the top 'k' treatments in terms of parent-perceived efficacy.

We present a novel implementation of multi-level association rule mining utilizing an external treatment ontology[1]. We discuss detailed experimental results in IAN dataset demonstrating the efficacy of our framework. Utilizing a host of measures including: (1) support, (2) confidence, (3) lift, (4) leverage, (5) conviction and, (6)chi-square test, we were able to prune the original rules generated to include only those with strong positive correlation.

It is important to note that the results discussed in this paper are based on the IAN data and are unique to this particular dataset. The results discovered here are not meant to replace the clinical scientist's results but are meant to be complementary to assist researchers to better assess the various treatments available using data mining algorithms.

The rest of the paper is organized as follows: We next discuss our approach in Section 4.2 and in Section 4.3 we describe our experimental results. In Section 4.4 we present our conclusions.

Approach

Medical data analysis is a challenging issue due to the inherently complex and critical nature of the data and the far reaching implications of the analysis in the medical domain. In this paper we propose a domain knowledge-guided framework which utilizes the well established foundation of association rule mining to generate strong association rules based on incorporating an external treatment ontology to guide the algorithmic process. Our approach comprises of the two following distinct components: (1) Association rule mining conducted on the entire treatment dataset and; (2) Multi-level association rule mining based on a dataset segmented according the treatment ontology. The multi-level association rule mining was conducted at two sub-levels: (a) First an analysis based on only categories and not the individual treatments was conducted in order to find the categories most associated with improvement; (b) Secondly an analysis for each

[1] All association rule mining was conducted using Weka. For more information on this data mining software go to www.weka.org.

category was conducted which included the treatments in order to find the top-performing treatments.

In addition to the more commonly used support and confidence, we incorporated lift, leverage and conviction measures in our analyses in order to detect negative correlation in the rules. Additionally we conducted a separate chi-square analysis in order to confirm the correlation as well as determine the strength of that correlation. Table 10 explains the interpretation of these measures.

Measure	Interpretation
Support	Minimum support was varied from .1 to .01 in order to pick up less commonly used treatments in the IAN dataset.
Confidence	Confidence was set very low .1 in order to pick up as many rules as possible.
Lift	< 1 - Negative correlation = 1 - Independent > 1 - Positive correlation The higher this value, the more likely that the rule is not just a random occurrence, but because of some relationship between them.
Leverage	< 0 - Negative correlation = 0 - Independent > 0 - Positive correlation
Conviction	< 1 - Negative correlation = 1 - Independent > 1 - Positive correlation Unlike lift, this measure is not symmetric which means it has no upper bound.
Chi-Square	The number itself denotes the strength of the correlation whereas a second heading is necessary to relate the direction of the correlation.

Table 10. Explanation of measures

Finally we take all the rules generated using our ensemble methods and perform a top-K analysis on them to find the strong association rules. We applied this framework on the IAN data to identify strong association rules between treatments for autism and the outcomes of the treatments with promising results. We describe the results in the following sub-sections: (1) The association rule mining on the entire dataset in Section 4.3.2 – Results with original IAN data; (2) The analysis of the category itself is presented in Section 4.3.3 - Results of IAN data as only ontology; (3) The treatment analysis incorporating the treatment ontology is presented in Section 4.3.4 – Results of IAN data with ontology and treatments and; (4) The top-k analysis can be found in Section 4.3.5 - Top-K analysis.

Experimental Results

Dataset

There were a number of data cleansing activities required in order to prepare the IAN dataset for analysis. This included determining the inclusion criteria; for example, we only included autistic children who had corresponding treatment data as well as SCQ (Social Communication Questionnaire) scores. As a result of these criteria, out of the 7,269 autistic children only 3,283 were included. A number of records were discarded due to data anomalies including treatments listed with no corresponding parent efficacy rating. In total, there were 14,351 treatment records with corresponding efficacy ratings.

There were three attributes used for this analysis: (1) Treatment name; (2) Treatment efficacy and; (3) Taxonomy. The treatment efficacy is rated by the parent using a 9-point Likert scale with four ratings for worsening, one for no change and four for improvement. During the course of our analysis, we found that consolidating the Likert scale was advantageous to the analysis particularly to garner enough support/confidence to produce meaningful rules. The consolidated Likert scale collapsed the four worsening categories into one category and similarly the four improvement categories were collapsed to one grouping. The no change category remained the same.

The ontology used was borrowed from the IAN website where a number of categories were presented to the users (parents of autistic children) as they would search for and enter treatment information. The categories for the taxonomy include: (1) Complementary and Alternative; (2) Educational and Behavioral; (3) Medications; (4) Physiological; (5) Special Diets and; (6) Vitamins and Supplements and (7) Top Ten.

Results with original IAN data

The top three treatments from the original dataset were not necessarily surprising since they are widely available for most autistic children in various settings (home, community, school etc.). The results of the statistical measures including chi-square emphatically confirm that ABA, Speech and Language and Occupational Therapy are indeed positively correlated with improvement (see Table 11).

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
ABA	0.93	1.11	0.01	2.27	61.06	Positive
Speech and Language	0.91	1.09	0.01	1.81	83.57	Positive
Occupational Therapy	0.88	1.04	0	1.29	16.66	Positive

Table 11. IAN Data

Results of IAN data as only ontology

When analyzing by ontology – three categories seem to have the strongest positive correlation with improvement – these include Educational and Behavioral, Top Ten and Medications. Top Ten is simply a category IAN used to compile the most commonly used treatments. Intuitively, associating the Top Ten category with improvement would seem judicious since most parents would probably not use treatments in such high numbers unless some improvement was noted. These results are listed in Table 12.

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Educational and Behavioral	0.88	1.04	0.01	1.3	77.37	Positive
Top Ten	0.88	1.04	0.01	1.29	59.4	Positive
Medications	0.86	1.01	0	1.06	1.52	Positive
Physiological	0.85	1	0	0.98	0.12	Independent
Vitamins and Supplements	0.71	0.84	-0.01	0.53	317.91	Negative

Table 12. IAN Dataset as Taxonomy

Results of IAN data with ontology and treatments

Sample 1: Complementary and Alternative

Originally, there were 5 treatments that were found to be associated with improvement. After looking at the statistical measures, it seems only Prayer is considered a strong rule with a lift value of 1.19, conviction of 2.55 and a very strong positive correlation. The rest of the treatments are considered weak rules with lift values below 0 (negative correlation), leverage values at 0 or below and conviction values below 1 (see Table 13).

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Prayer	0.92	1.19	0.02	2.55	268.71	Positive
Weighted Blanket	0.74	0.96	-0.01	0.88	1.18	Negative
Chelation	0.74	0.96	0	0.84	0.5	Independent
Music Therapy	0.73	0.95	-0.01	0.99	0.5	Independent
Equine Therapy	0.72	0.93	-0.01	0.8	2.1	Negative

Table 13. Complementary and Alternative

Sample 2: Educational and Behavioral

Originally, 6 treatments were found to be associated with improvement. In this analysis, ABA, Speech and Language Therapy and Visual Schedules are confirmed by their statistical measures (see results in Table 14). PECS and Social Stories were negatively correlated – Social Skills was not present in the lift analysis but was included for the chi-square analysis regardless.

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
ABA	0.93	1.05	0.01	1.66	24.02	Positive
Speech and Language	0.91	1.03	0.01	1.32	21.89	Positive
Visual Schedules	0.9	1.02	0	1.16	1.32	Positive
PECS	0.85	0.96	0	1	6.19	Negative
Social Stories	0.83	0.94	0	0.66	12.77	Negative
Social Skills	0.79	N/A	N/A	N/A	46.78	Negative

Table 14. Educational and Behavioral

Sample 3: Medications

Clonidine was confirmed as a positively correlated rule. Risperidone, on the other hand, had a lift measure of .99 and conviction of .95 which suggests a negative correlation and a chi-square value indicating independence (Results listed in Table 15). This is quite surprising since risperidone is one medication for autism that has shown efficacy in clinical trials (Aman, Arnold et al. 2005).

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Clonidine	0.92	1.08	0.01	1.73	6.61	Positive
Risperdal	0.85	0.99	0	0.95	0.06	Independent

Table 15. Medications

Sample 4: Physiological

Of the four treatments that came up with strong confidence, three remain as strong rules (sensory integration therapy, occupational therapy and physical therapy). Weighted blanket (vest) had a lift of .88, leverage of -.01 and conviction of .59 indicating poor (negative) correlation. These results are presented in Table 16.

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Sensory Integration Therapy	0.89	1.05	0.01	1.36	6.49	Positive
Occupational Therapy	0.88	1.04	0.02	1.25	20.72	Positive
Physical Therapy	0.87	1.03	0	1.14	1.52	Positive
Weighted Blanket or Vest	0.74	0.88	-0.01	0.59	20.67	Negative

Table 16. Physiological

Sample 5: Special Diets

There were conflicting results with the inclusion of the other statistical measures for special diets (see Table 17). The casein-free, milk-free and gluten-free diets which were shown to be effective in the previous analysis – are all showing negative correlation in the lift and leverage measures. Also, now there are new rules generated stating there is a correlation between “no change” and the gluten-free and casein-free diet. The confidence levels are not high for these rules and the lift/conviction measures are signifying independence.

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Gluten-free diet (No Change)	0.2	1.1	0.01	1.02	0.76	Independent
Casein-free diet (No Change)	0.18	1.02	0	1	0.03	Independent
Casein-free diet (Improvement)	0.81	0.99	0	0.95	0.09	Independent
Dairy-free (milk-free) (Improvement)	0.81	0.99	0	0.92	0.08	Independent
Gluten-free (Improvement)	0.8	0.98	-0.01	0.9	0.64	Independent

Table 17. Special Diets

Sample 6: Vitamins and Supplements

Omega-3 fatty acids were shown to be negatively correlated with improvement. However, melatonin is by far the strongest treatment in this category with a strong positive correlation. Probiotics is also a strong treatment in terms of the statistical measures. These results are depicted in Table 18.

Treatment	Confidence	Lift	Leverage	Conviction	X^2	Correlation
Melatonin	0.94	1.32	0.02	4.53	62.84	Positive
Probiotics	0.76	1.07	0	1.17	2.11	Positive
Omega-3	0.63	0.89	-0.01	0.77	6.43	Negative

Table 18. Vitamins and Supplements

Top-K analysis

Table 19 lists the top 10 treatments as well as the associated measures of confidence and tests of correlation. ABA, Occupational therapy and Speech and Language are listed twice since they showed up in both analyses – TD refers to “total dataset” and SD to “segmented dataset.” The following criterion was used to include treatments in the top ‘k’:

1. Confidence > .70

2. Two of the three statistical measures (lift, leverage and conviction) had to indicate positive correlation. For lift and conviction this meant for any value above 1 and for leverage any value above 0.

3. A positive correlation for the chi-square test.

This analysis of the top-analysis does not indicate treatment efficacy, since that can only be determined in a properly controlled clinical study. However, out of the 500 treatments that are present in the IAN dataset, this top 10 list indicates those treatments most like to be effectual and would be good candidates for clinical study.

Treatment	Confidence	Lift	Leverage	Conviction	Chi-Square
Melatonin	0.94	1.32	0.02	4.53	62.84
ABA - TD	0.93	1.11	0.01	2.27	61.06
ABA – SD	0.93	1.05	0.01	1.66	24.02
Prayer	0.92	1.19	0.02	2.55	268.71
Clonidine	0.92	1.08	0.01	1.73	6.61
Speech and Language - TD	0.91	1.09	0.01	1.81	83.57
Speech and Language - SD	0.91	1.03	0.01	1.32	21.89
Visual Schedules	0.9	1.02	0	1.16	1.32
Sensory Integration Therapy	0.89	1.05	0.01	1.36	6.49
Occupational Therapy - TD	0.88	1.04	0.02	1.25	20.72
Occupational Therapy - SD	0.88	1.04	0	1.29	16.66
Physical Therapy	0.87	1.03	0	1.14	1.52
Probiotics	0.76	1.07	0	1.17	2.11

Table 19. Top-K Analysis

Conclusion

Multi-level association rule mining is an effective tool in allowing domain knowledge to guide model development. Through the use of a treatment ontology, we were able to extract more meaningful rules from the IAN dataset. Subsequent analyses of measures such as lift, leverage and a separate chi-square test pruned the list of treatments to include only those that are positively correlated. As a result, a top-k list of treatments for autism was presented.