Ryan Littleton

Data Mining DSC- 607

Course Project- Final

**Introduction**:

In the continuing effort to deal with the problems of child abuse and neglect, one specific area that is always in need of additional resources and attention is that of preventative measures. In the World Health Organization’s *World Report on Violence and Health,* many suggestions are defined for prevention from action by law enforcement, health service providers, educators and communities. (Krug et all, 2015) In many cases, these efforts are for purposes of identifying and intervening, but in many others, these are for preventative measures, such as community education and initiatives to rectify risk factors. Two of the challenges in the report that are of particular interest to this proposal are a lack of available and interpretable data in some areas and also the need for more research into risk factors for abuse. Many are already known, however quantitatively their extent, importance and also the existence of other factors are much less clear. The use of data mining techniques to better address these issues is the core motivation behind this study.

It is of ethical importance to note that in no case would any of this mined information be used for identifying specific individuals as victims or perpetrators, but rather for identifying areas and individuals that are at higher risk to divert resources to that could be used in the manners of intervention mentioned initially, or to more effectively identify different categories among those vulnerable to abuse and neglect. The results can also have the dual purpose of assisting future research into risk factors. Knowledge of what factors comprise the higher risk clusters can give an indicator of areas for future research into risk factors, which would help facilitate taking a more preventative approach to child abuse and neglect.

This area is particularly suited to data mining in particular due to several factors. First, the volume of demographic data kept by various institutions in this nature is very high, making it prohibitively difficult to manually analyze. Second, this area of research has a demonstrated need for both useful identification methods and avenues for further investigation, both of which are assisted by the proposed system. Finally, the data available contained is sufficiently diverse to be examined in multiple ways and to also be suited to further research.

A final note must be made on the ethics of this investigation. Because it is such a sensitive topic, is important to note that this system should not be used to replace in person investigation methods or for pejorative purposes for any individuals. It can be used as a tool to enhance research or investigation, but if it ever replaced it, the risk of ethical violations against individuals would be much too high. That being said, the results of this study could still be extremely useful outside those dangerous use cases, and it is the author’s opinion that under ethical use, it could greatly assist efforts to both focus the preventative efforts as regards child abuse and neglect while also opening new avenues for further research.

**Analysis method**

In efforts to address the problems of child abuse and neglect, it is clear that efforts in prevention are equally or perhaps more important than those of enforcement. The *Child Maltreatment* report from the Department of Health and Human Services Administration for Children and Families makes this clear with the statement, “Child protective services (CPS) agencies promote children’s safety and well-being with a broad range of prevention activities and by providing services to children who were maltreated or are at-risk of maltreatment” (US Department of Health and Human Services, 2017). It is, then, of utmost importance to keep minding effective management of information regarding risk factors, preventative measures and what areas and individuals are most at need of intervention. For instance, it is well established that individuals whose parents were abused are at a higher risk of abuse and that this risk can be mitigated with services that assist the formerly abused parents “…by moving beyond parenting programs alone to incorporate the broader community and societal contexts that can help ensure the conditions for good health and well-being.” (Merrick and Guinn, 2018) It is with this understanding that the proposal to closer investigate the existing data on child abuse and neglect in order to better evaluate risk factors and generate new indicators for effective research into prevention methods is proposed.

To that purpose, the *Fourth National Incidence Study of Child Abuse and Neglect (NIS–4)* was selected for the application of selected data mining techniques. (Sedlak, 2006) This dataset is the most recent publically available such study and contains several features making it suited to such research. It differs from many similar studies in that it contains not only contains data reported by CPS agencies, but also data on maltreatment reported by professionals that was either screened out by CPS agencies or was not reported to them. This makes the dataset more complete for preventative purposes, where the standard of being able to be legally prosecuted is an overly stringent screening requirement. Another feature is that it contains in parallel data on subjects that meet the harm standard as well as those who meet a less stringent endangerment standard for inclusion into the data. The dataset contains 12,408 records of children countable by this endangerment standard, as well as 286 drug effected newborns, who were included but are not considered countable by the abuse standards for the study. There are 1844 variables, many of which are binary, and many of which are not part of the data of the study but contain some other information about the data.

Before discussing screening techniques, it must be noted that all data preparation tasks and algorithm implementation was performed using R version 3.5.1. In initial data preparation, many of the extra variables mentioned above were attended to. First, several are weights that are considered mandatory by the study designers. They are such because of the study design containing some known over or underrepresentation of children with certain attributes. For instance, as noted in the descriptions in the metadata document, children who died as a result of abuse are overrepresented due to the very high reporting rates of such cases. As such, these weights must first be applied, then removed from the variables for analysis. Several more variables are countability variables or variables related to the agencies who reported the abuse. These variables were screened from the investigation, because while they contain valuable data, they are outside the scope of this particular study. Upon close examination, many variables were found to contain redundant information, as they were either binarized versions of the initial categorical data or they were aggregations of other variables. Due to the nature of the algorithms used, which will be discussed below, the binarizations of other variables were retained and the others were screened out. This was accomplished through regular expressions once it was discovered that the dataset had a particular organization and regularities with regards to the naming conventions of variables in these categories. Outliers were also screened for, using the local outlier factor algorithm, a density based screening algorithm. This one was chosen rather than traditional statistical measures for outliers due to the large number of variables making such methods impractical. It became clear upon employing this algorithm that the most extreme outliers in the dataset existed for a systemic reason, namely the inclusion of the drug affected newborns who did not meet the harm or the endangerment standards for countability. Due to this factor, they were removed from the analysis, as this revealed them to be outside scope of the study. Missing values were checked for, and found to exist in high numbers throughout the dataset. These were also found to mostly exist for a systemic reason. Several of the binary variables were form questions that included a yes or no question that was only answered if an individual fit some other metric. For example, a question on whether CPS had investigated maltreatment in a particular case had a field for “yes”, for “no” and “CPS has no record of child” which was coded NA. (Sedlak, 2006) Because these NA cases contained in almost all cases a meaning that for the purposes of this study is equivalent to no, they were recoded as “no” or 0 values.

The first technique employed was association analysis. Association analysis reveals rules that represents connections between different points in the data. With this technique, the variables that have positive values in conjunction with each other frequently enough will be deemed a sufficiently associated with each other to form association rules. The ruleset thus generated was pruned through use of an apriori algorithm to filter those rules that are not well supported enough by the dataset to be interesting to the analysis. This technique was especially well suited to this dataset given the binary status of such a large number of variables, giving them the ability to be simulated in the algorithm as TRUE/FALSE pair items.

This approach to mining this dataset gives several advantages relevant to the goals of this study. First, the ability to connect different qualities of abuse victims along with the types of abuse associated with these groups can be used in future research as part of a road map to how preventative measures can be applied, including as advisory data to social workers and researchers on community and personal assistance operations that might be offered to parents and caretakers of children whose households have similar qualities. The associations revealed can also be used to inform future qualitative analysis. Knowing which features are associated with each other can give guidance for determining how groupings might be classified.

The second data mining method used in this analysis is clustering using the a fuzzy clustering implementation of K means, which uses updating of centroids to find optimum clustering of a specified number of groupings of the dataset.

Clustering is useful in this case, as it can reveal what sets of qualities appear together in this data. By revealing how the subjects cluster together around sets of attributes, different profiles of types of victims of abuse can be revealed, which has the benefit of allowing a classification to be developed. Such a classification might be used to identify types of interventions that might work as preventative measures in communities where qualities revealed in the clustering are found in higher amounts. As such, this technique will offer useful information that may not be a predictive element in and of itself but can offer valuable data to future predictive work. Furthermore, a knowledge of how these features cluster can again be useful to future efforts to gather data in the ways that are the most beneficial to the task of prevention. It should be noted that initially, this study used an implementation of the Jarvis-Patrick algorithm clustering instead, chosen due to its ability to effectively handle high dimensionality in data and different sized clusters, but the optimization of that algorithm proved to be prohibitive, leading to results with too many clusters to be interpretable, so fuzzy clustering, or c-means, was instead chosen to allow for a more useful result, despite its limitations.

**Results**

After applying the apriori algorithm to the frequent item ruleset generated from the dataset, using support threshold of .1 and a confidence threshold of .9, there were still clearly more rules in the itemset than could effectively be evaluated in a meaningful way, so several different pruned versions of the itemset were generated with the same parameters, only specifying the fields in the “rhs” side of the rules to have specific sets of values. Redundancies were also removed. The four rulesets generated in this way were a ruleset with all rules resulting in an item having to do with the endangerment standard, one with all involving the harm standard, one with all cases having to do with emotional abuse or neglect and one with physical abuse or neglect. The ruleset for physical abuse and neglect only had two rules after pruning, so another version was generated with a lowered support threshold as well. The same was also attempted with sexual abuse and neglect, but no rules were generated that met any reasonable pruning thresholds. Beyond these objective measures of interestingness, there are several subjective measures that will bear importance in the interpretation of these rulesets. Associations that might reveal useful information relevant to the stated goals of this study are of particular importance, as well as avoiding a semantic redundancy that might not have been caught by algorithmic redundancy measures. Given these metrics, the tables of rules for emotional abuse and neglect as well as for the harm standard contained the highest amount of subjectively interesting rules, which will be addressed in the interpretation section.

For the cluster analysis, the dataset was first converted to a difference matrix and then the fuzzy clustering algorithm was run. An elbow plot was made for clusters between k=1 and k=10, revealing the optimal number of clusters to be 2. The clustering coefficient for the algorithm with k=2 was almost zero, indicating the fuzzy clustering attempt was very fuzzy with little separation. This likely indicates that even with the optimization, the clustering will not contain very much information on the categorization of the data.

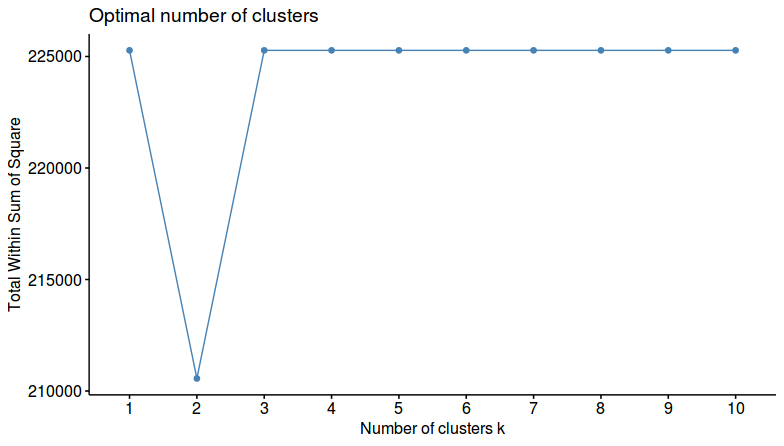


Figure 1: Elbow Graph for Cluster Analysis

**Interpretation of Results**

Of the two methods used in this study, the association analysis yielded more immediately usable results. As previously mentioned, the emotional abuse and neglect as well as the harm metric rulesets contained the most subjectively interesting association rules of all the possibilities attempted.

Starting then, with the ruleset regarding emotional abuse and neglect, one thing that stands out is that all the rules after pruning specifically reference emotional neglect rather than abuse. Two had confidence of 1, indicating complete association, but these rules turned out not to be particularly interesting as they contained a hidden semantic redundancy. They were associations between age categories over 26 and over 35 for perpetrators who committed emotional neglect. Looking at some of the rules of lower, but still significant confidence and support in this table, then, is seen an association between perpetrator drug use being considered a factor and emotional neglect. This can be seen, for example, in the rule: {ANEGE,PRPDRUGE1} => {ENE}, which has a confidence of .81. This indicates an association between drug involvement and the neglect endangerment standard with emotional neglect. What made this particularly interesting is that none of the other association tables included drug use as involved in a strong association with their abuse standards. Given its presence in this table and not others, it may then be concluded that drug use is more likely to be a factor in emotional neglect than in other forms of abuse and neglect, or that it is more likely to be recorded as a factor in those cases. Below is the graph for these sets of standards.

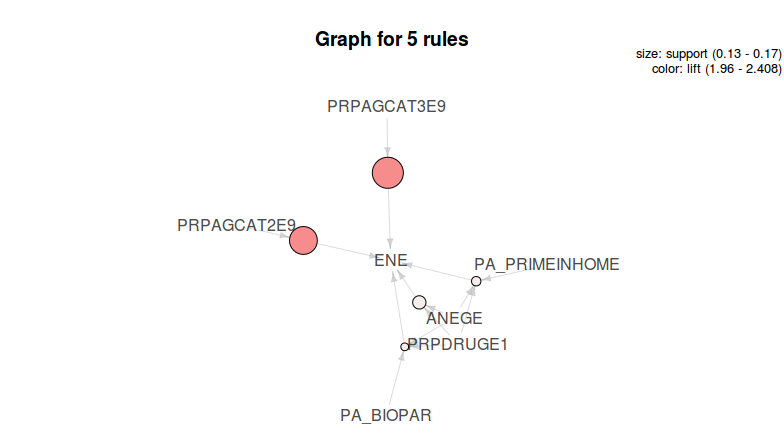


Figure 2: Graph of Rules Regarding Emotional Abuse and Neglect

With regards to rules that meet the harm rather than the endangerment standard, the most frequent of the labels filtered by rhs to feature in our ruleset is PAH, or any physical standard of harm or abuse was met. While this standard itself is very general, the noteworthy item to mention is that for multiples of these rules, one of the items that was involved in this rule was an item that indicated the perpetrator of abuse was the biological parent. The confidence for these rules were very high, and no other relationship to the perpetrator was noted in these rules, so a conclusion can be drawn that the biological parents are individuals of very important concern when it comes to prevention measures. An example of these sorts of rules is {AABUSEH,PAE,PA\_BIOPAR} => {PAH}, with a confidence of .962. AABUSEH indicates the presence of some abuse meeting the harm standard, PAE indicates physical endangerment standard met and PA\_BIOPAR indicates a biological parent was the perpetrator. As mentioned, multiple rules of these sorts of combination contained an indicator of biological parenthood of perpetrators, but none for other relationships. This does not indicate the perpetrator is never of another relationship, only that only those who are biological parents met the pruning standard. The only item in this itemset that is not a redundancy that screening failed for was {AABUSEH,PAE,PA\_PASTEMPUNK} => {PAH}, with confidence of .965 where the first two codes are as specified above, but the third code, PA\_PASTEMPUNK indicates the past employment of the perpetrator is unknown. This is interesting as one of the goals of this study is finding areas where further research or efforts could be indicated, and the frequent item specified is of unknown employment, which could indicate a failure to report or gather this information. This is not simply a case of the perpetrator having no past employment, as that information is captured in another variable in the dataset. Below is the graph for the ruleset screened for the harm standard.

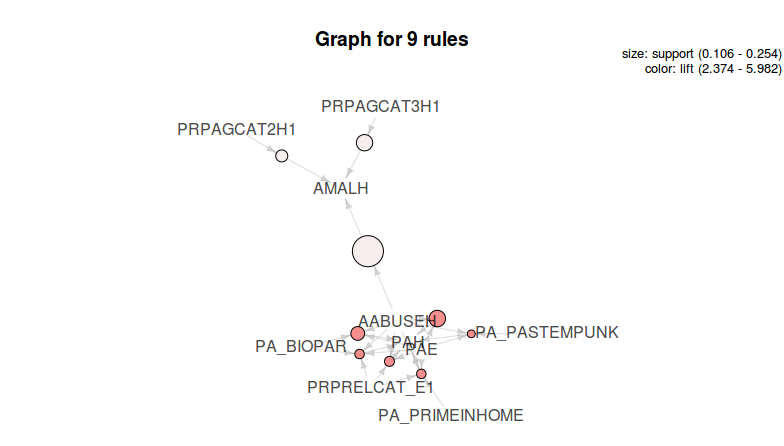


Figure 3: Graph of Rules Regarding Harm Standard

The last area with results to be interpreted is the fuzzy clustering. As mentioned, the cluster quality was not very clear, even after optimization steps were taken. This may indicate a lack of clear categories in this dataset or perhaps that a different algorithm could have been more effective. Optimizing the Jarvis Patrick algorithm turned out to be too computationally expensive, but that does not necessarily indicate that a different one might not be more effective. In any event, the results of this clustering are unlikely to yield any meaningful insights as they currently stand, and the recommendation from this study is to leave those efforts for a subsequent analysis.

**Limitations of the analysis**

There are several points of limitation that must be noted regarding this analysis. First and foremost, the ethical caveats of the earlier explanations must be reiterated. The potential for misuse is too high for a topic as important as this one not to apply an extra layer of human judgement and using this data for any effort that might be pejorative would be a mistake. A second limitation of this analysis is that the child abuse data by necessity only includes children who have met some standard of victimization. With the initial goal of assisting in prevention, not being able to screen for what factors are associated with victimization verses non-victimization puts a hard limit on the boundaries of predictive power of any efforts taken here. Last, the multidimensionality of the dataset presents a large number of opportunities for meaningful generalizations to be buried in the analysis and limits the potential usefulness of some algorithms, as is quite evident from the results of clustering. All of these issues connect to what areas might be suggested for subsequent analysis.

**Suggestions for a subsequent analysis**

This analysis being a limited analysis on a very large and also important issue leaves many potential avenues for further study. One area that would be beneficial would be to acquire more data, especially labeled data, so as to repeat the attempts at categorization with supervised categorization rather than clustering, which may prove more fruitful. Another area would be to find data that takes locations into account so as to identify areas of greatest need for services and prevention measures.

**Conclusion**

The association analysis portion of this study provided some insights into what aspects of a victim’s environment are connected with different standards of abuse and abuse type, but the attempt at using a clustering algorithm turned out to be too complicated for the resources available in the case of the Jarvis-Patrick algorithm and to yield suboptimal clustering in the case of the simpler c-means algorithm. While these results are limited, they do indicate several avenues for future study, and at least in the case of the association analysis, give a few indicators of what steps might be taken by people providing services in high risk areas or conducting data collection regarding abuse.

**References**

Krug, E., Dahlberg, L., Mercy, J., Zwi, A., & Lozano, R. (Eds.). (2015, February 02). World report on violence and health. Retrieved September 23, 2018, from http://www.who.int/violence\_injury\_prevention/violence/world\_report/en/

Merrick, M. T., & Guinn, A. S. (2018). Child Abuse and Neglect: Breaking the Intergenerational Link. American Journal of Public Health, 108(9), 1117-1118. doi:10.2105/ajph.2018.304636

Sedlak, A. (2006). *Fourth National Incidence Study of Child Abuse and Neglect (NIS–4)*. Retrieved October 7, 2018, from <https://www.ndacan.cornell.edu/datasets/dataset-details.cfm?ID=147>.

U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children’s Bureau. (2017).

Child Maltreatment 2015