George Hodulik

EECS 435

Proposal for: Finding Relevant Overlapping Subspace Clusters in Social Justice Sexuality Project using ROCAT algorithm

1. Project Idea

For my EECS 435 Data Mining project, I will implement the ROCAT algorithm presented in the paper, "Xiao He, Jing Feng, Bettina Konte, Son T. Mai, Claudia Plant: Relevant overlapping subspace clusters on categorical data. KDD 2014, 213-222," [2] on the public data set from "Social Justice Sexuality Project: 2010 National Survey, including Puerto Rico (ICPSR 34363)." [1]

2. Brief and simplified description of ROCAT algorithm

The ROCAT algorithm efficiently finds relevant subspace clusters in categorical data. The algorithm uses Shannon entropy to calculate the Minimal Description Length of a categorical data set with regard to specific subspace clusters. The algorithm iteratively finds the best "pure" subspaces (a subspace in which all attributes have the same value), and checks to see if each new subspace cluster decreases the MDL of the overall data set; if it does, it will later try to expand that subspace. This part of the algorithm continues until there are no new subspace clusters to expand (further clustering does not decrease MDL) [2].

Then, the algorithm examines the subspace clusters, and combines/splits those that overlap (shown below) such that the MDL of the data set is lowest.

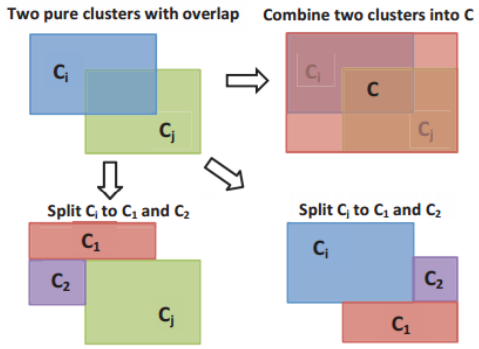

Figure 3: 4 processing candidates in Combining phase.

[2]

Lastly, there is some post-processing to remove some redundancies that may have occurred [2].

3. Survey of Related Work

While there is much research on clustering numerical data, there are not very many studies that focus on categorical data. Some subspace cluster algorithms for numerical data, like CLIQUE, can be implemented to find subspace clusters in categorical data. However most algorithms that can be implemented for categorical data have flaws: LIMBO does not take into account model complexity; CACTUS, SUBCAD, and LIMBO cannot find overlapping subclusters; CLICKS and CLIQUE often includes too many redundant clusters. Some numerical algorithms that try to fix these problems, like STATPC and RESCU, cannot be applied to categorical data and require parameters [2].

Pattern mining algorithms can usually be applied to categorical data because frequent itemsets can be analogous to subspaces. However, translatable pattern mining algorithms typically need fault-tolerant itemsets, and/or require input parameters like minimum support [2].

Algorithms which do support parameter-free subspace clustering of categorical data, like DHCC and AT-DC, are greatly affected by outliers [2].

ROCAT consistently runs in a similar amount of time (or significantly better) than comparable algorithms, and produces better results [2].

4. Data Set

After searching for large data sets with mostly categorical attributes, I have found the survey results from the Social Justice Sexuality Project 2010 National Survery, an ideal data set for the ROCAT algorithm. This data set contains the survey results of "one of the largest national surveys of Black, Latina/o, Asian and Pacific Islander, and multiracial lesbian, gay, bisexual, and transgender (LGBT) people [1]." The survey asks subjects from all 50 states, Washington D.C., and Puerto Rico socio-political questions with special regard to five factors: racial and sexual identity, spirituality and religion, mental and physical health, family formations and dynamics, and civic and community engagement [1]. This survey is ideal for the ROCAT algorithm because it is almost entirely categorical, and has a size in which the algorithm should run smoothly (but not trivially easily).

Shown below, the ROCAT algorithm was tested at two extremes (other tests not shown here): number of objects, in which a data set containing up to 50000 objects with 52 attributes took a few minutes (estimating from observing graph) to run, and dimensionality, in which a data set containing 960 objects with up to 200 attributes took less than 20 seconds to run [2].

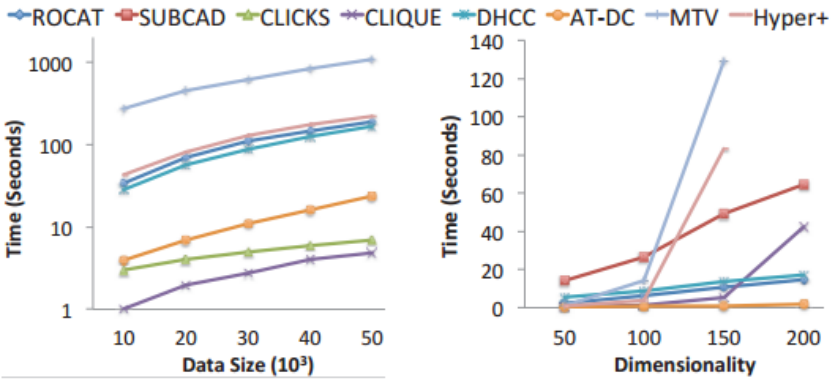

Figure 6: Scalability of ROCAT and comparisons.

[2]

This survey contains almost 5000 data rows with over 100 attributes. Based on the above data, this data set should be a good size for this algorithm: large enough to test the algorithm, but not trivially small.

5. Implementation

The data set is downloadable in a tab delimited format. I will import the data set to local storage in an SQL table. For implementing the ROCAT algorithm, I will use either Python or Java, as these are languages I have experience in dealing with SQL. The algorithm should not require any highly specialized libraries: it calls for matrices, queues, priority queues, and easy calculation of Shannon entropy, all of which can be implemented in either Java or Python.

The data set will require some preprocessing. Firstly, several attributes were blanked for confidentiality and replaced with a new, less specific, categorical attribute; the blanked questions will need to be removed. There are also some attributes that primarily have the same responses, and these I will also likely remove. For data objects which are missing responses (the subject refused to answer), I will most likely treat refusing to respond as its own category; if it is reasonable for certain questions, I may replace empty responses with the average of that response.

Furthermore, there are some, but not many, quantitative attributes. For these, I plan to map them to categories corresponding to ranges. Most, if not all of these, have obvious mappings: for example, there are ratings questions which ask the subject to rate from 1 to 6 (or 10) how much they agree with a statement - these I may map to simply "Agree (1-2)," "Neutral (3-4)," and "Disagree (5-6)."

It may be necessary for me to take into account shift relationships for ratings question because of rating inflation discussed in class (some people rate things overall higher/lower than others). However, I think mapping ratings to more broad categories, as already described, should adequately resolve this problem.

Finally, while most of these responses are in the form of integers representing a category, there is the question with a text response asking what state the subject is from. I will probably map the states to integers to make them easier to process.

6. Anticipated Results

I believe the resulting subspace clusters found by implementing ROCAT on the Social Justice Sexuality Project study may have very interesting, revealing results. The algorithm may find that a set of questions has primarily a much smaller set of distinct responses, which could have very strong implications.

More specifically, I think that any subspace that is age-specific could have very intriguing implications. In this survey, age is split into 3 groups: 18-24, 25-49, and 50+ (exact age was blanked for confidentiality). This categorical grouping makes finding generational trends or distinctions convenient to find. Responses to certain questions could have very revealing findings if there is found to be generational trends in responses. Some of these interesting questions include:

"What are the three (3) most important issues facing you?"

"In your opinion, what are three (3) most important issues facing LGBT communities of color in the U.S.?"

"[Rate how much you agree:] Homophobia is a problem in my neighborhood"

"Do you think that same-sex marriage should be legalized?"

There are also several other socio-political factors in the survey, including race/ethnicity, education, religion, gender identity, etc; trends found with respect to any of these factors could also have very powerful implications.

7. Possible Further Implementations (time permitting)

In order to find trends in the data set that may be specific to certain socio-political factors, I may try running the algorithm on a projection of the data set which only contains certain socio-political factors. For instance, I may want to see if there is an overall trend with respect to gender identity, an attribute that should be independent of all the other attributes.

Lastly, I may also implement other algorithms that were compared against ROCAT. The paper compares ROCAT to SUBCAD, CLICKS, CLIQUE, DHCC, AT-DC, Tiling, MTV, and Hyper+; I may test some of these to compare their performance ROCAT.

8. What I expect to submit

At the end of the semester, I will submit a report describing my project in detail. I will describe my process of implementing the project, describing setbacks I may have had, as well as a description of my final implementation. In my description of my final implementation, I will explain the ROCAT algorithm in detail, and how I translated the pseudocode to a real implementation.

My report will also contain any results I found about the data set from implementing ROCAT. I will discuss the significance of the results. I will also compare the results and performance of running ROCAT to any other algorithm(s) I may implement.

9. References

[1] Battle, Juan, Antonio Jay Pastrana, and Jessie Daniels. Social Justice Sexuality Project: 2010 National Survey, including Puerto Rico. ICPSR34363-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-08-09. http://doi.org/10.3886/ICPSR34363.v1

[2] Xiao He, Jing Feng, Bettina Konte, Son T. Mai, Claudia Plant: Relevant overlapping subspace clusters on categorical data. KDD 2014, 213-222. http://dl.acm.org/citation.cfm?id=2623652