Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining

1. What is the problem that the paper wants to solve? Why is it difficult (related works)?
   - Problem definition: Given a data matrix, find association rules which define sets of predictive entries or more precisely which given some attributes can be used to derive another set of attributes
   - Related to prediction problems in statistics and machine learning but on bigger datasets, thus needs to be fast enough to minimize the time required to find these association rules
   - Traditional approaches require multiple passes over the data or large amounts of memory
2. What is the solution? What is the main idea?
   - Instead of association rules derive ratio rules in the form $a : b : c$ for columns a, b, c, i.e. customers spend $1 : 2 : 5$ dollars on bread : milk : butter by eigensystem analysis
   - By computation of the top-k eigenvectors and eigenvalues of the covariance matrix identify the direction of maximum variance and then incrementally orthogonal directions of maximum variance. These eigenvalues then give the ratio rules.
   - Authors introduce root-mean-square error as a measure of “goodness" to assess the derived association rules
3. What is the result?
   - Ratio rules allow for extrapolation and prediction, give compact representations of linear correlations, and are easily implemented
   - Scale well for large datasets as they are fast to compute, growing linearly on the largest dimension of the matrix (typically rows)
   - Authors provide a new metric in order to evaluate their approach
4. What is the main novelty that enabled the solution?
   - The main novelty is using eigensystem analysis to derive the top-k ratio rules allowing us to determine them in a single pass over the dataset
5. What are the good aspects of the paper? Did you learn something from the paper?
   - Computes ratio rules in a single pass over the data, small memory requirements (as opposed to traditional association rule mining methods)
   - Authors use concrete example (customer $\times$ product matrix) to better visualize the idea
6. What is the impact of the paper?
   - Provided a completely new type of rules that capture linear correlations in data as well as a new measure to evaluate the “guessing error" of ratio rules and similar predictive rules
7. Are there weaknesses/missing parts in the paper? How can you improve it?
   - Only better suited for the problem than quantitative association rules if data is linearly correlated. For clustered data association rules will provide better accuracy.
8. How can you extend the paper?
   - Explore how ratio rules perform on non-linearly correlated datasets. Similar to ratio rules, how can we find rules for non-linear data?
9. How can you apply the technique to other data/problems?
   - Sometimes approximation is sufficient: use top-k values for an estimate of sufficient accuracy but much better runtime speed (in this case eigenvalues but applies to other algorithms as well)