Detecting Simpson's Paradox

Chenguang Xu
University of Oklahoma
Norman, Oklahoma
chguxu@ou.edu

Sarah M. Brown
University of California, Berkeley
Berkeley, California
sarahmbrown@berkeley.edu

Christan Grant
University of Oklahoma
Norman, Oklahoma
cgrant@ou.edu

Abstract

Simpson's paradox is the phenomenon that a trend of an association in the whole population reverses within the subpopulations defined by a categorical variable. Detecting Simpson's paradox indicates surprising and interesting patterns of the data set for the user. It is generally discussed in terms of binary variables, but studies for the exploration of it for continuous variables are relatively rare. This paper describes a method to discover Simpson's paradox for the trend of the pair of continuous variables. Correlation coefficient is used to indicate the association between a pair of continuous variables. We use categorical variables to partition the whole data set into groups. Our algorithm's goal is to find the sign reversal between the coefficient correlations measured in the group relative to the original entire data. We show that our approach detects cases in real data sets as well as synthetic data sets, and demonstrate that our approach can uncover the hidden surprising pattern by detecting occurrences of Simpson's paradox. This paper also proposes an approach that exploits sampled data for early Simpson's paradox detection. We show the running time for the algorithm by examining through the combination of different conditions.

Introduction

Discovering insights from data is a crucial aspect of data science. The public and overseers are increasingly scrutinized because important data sets contain many surprising results that are left unexplained or unexplored (Doshi-Velez et al., 2017). Simpson's paradox is one of the most well-studied surprising trends in data. Developing an automated method for the detection of this paradox will help industries scrutinize the data sets that effect everyday life.

Simpsons paradox is the reversal of the relationship between a pair of variables when conditioned on a third variable. The phenomenon may occur within all of the subgroups or for some. We categorize cases of Simpson's paradox into two cases, based on the type of trend to be reversed: *classification*, when the trend is the relative rates of a binary outcome in two groups and *regression*, when the trend is based on the sign of a correlation between two variables.

Figure 1 shows a synthetic example of Simpson's paradox. A black-dotted line shows a regression line over the full data set, indicating positive correlation. However, two subsets of the data sets, shown as red-circles and blue-x, individually have negatively sloped regression lines. Conditioning on this symbol type, we see an opposite relation between the two attributes. If analysts reach conclusions based on data that has such disparities, lives and livelihoods may be affected.

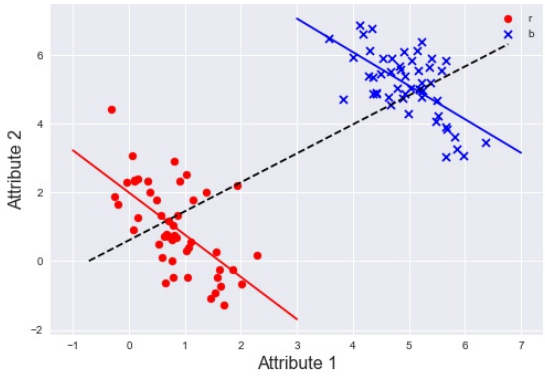


Figure 1: Simpsons' paradox occurring in an example data set. Full weight trendlines are for subgroups and the dashed line is the population trend

While visualizations can help analysts discover the existence of this paradox, as data sets are increasing in dimensionality and growing in size, we can no longer rely on visual inspection. This motivates development of robust automated detection of Simpson's paradox.

In this paper, we propose an algorithm to detect Simpson's paradox for the *regression* case and demonstrate its empirical utility on three data sets. We then show the performance of the algorithm over samples of full data sets and large data sets.

Background

A famous example of Simpson's paradox is in relative rates of graduate admission by gender that reverses when departments are considered individually (Bickel et al., 1975). Simspon's Paradox has also been observed in drug dosage

to outcome analyses where both genders show negative trends, but gender differences drive a populaiton-wide positive trend (Kievit et al., 2013).

The phenomenon is well-studied in statistics (Pavlides and Perlman, 2009; Chen, Bengtsson, and Ho, 2009; Lerman, 2017) and through the lens of causality (Pearl, 2011; Hernán, Clayton, and Keiding, 2011; Arah, 2008). The association between two variables is considered and studied in Alin (2010), and the causal-theoretic view point is examined. Bandyoapdhyay et al. (2011) also argue the causal account of the Simpson's paradox and provide another perspective comparing with classic work by Blyth (1972) for the logic of Simpson's paradox.

Generic detection of Simpson's Paradox has been done with respect to available discrete attributes in the data (Guo, Binnig, and Kraska, 2017; Freitas, 1998) including with ranking occurences (Fabris and Freitas, 2000) and clusters discovered within the data Kievit et al. (2013). Techniques for visually detecting Simpson's paradox (Armstrong and Wattenberg, 2014) and more general surprising results (Rücker and Schumacher, 2008) also exist. Simpson's paradox's impact on learning has been studied with respect to reliability of association rules (Froelich, 2013).

Methodology

Simpson's paradox has been studied in two main forms: relative rates and linear trends. We focus on the latter and use linear correlation to measure a trend between two variables.

Detecting Simpson's Paradox

We formally describe the algorithm in Algorithm 1. Given a data set, we assign each column to one of two lists: (1) *group-by attributes* for conditioning over (integer or nonnumerical columns) ; (2) *candidate attributes* for computing the relationships (continuous valued columns).

For a data set with d candidate attributes we compute the $d \times d$ matrix of correlation coefficients. Next, we partition the data set by conditioning on each of the C group-by attributes and compute an additional $d \times d$ correlation matrix for each of the k_c values of attribute c. In total, we compute $\sum_{c=1}^{C} k_c + 1$ correlation matrices of size $d \times d$. An example from our synthetic data set is shown as Table 1.

Finally, for each pair of candidate attributes (the upper halves of the correlation matrices), we compare the sign in each of the $\sum_{c=1}^{C} k_c$ subgroup-level matrices to the sign of that pair in the whole data. For each sign reversal found, we record the correlation of whole population(**allCorr**), reversed correlation value(**revCorr**), the pair continuous attributes (**attr1**$= a_1$, **attr2**$= a_2$) that exhibit the reversal, the categorical attribute(**catAttr**, c), and the **subgroup** value, s. The output of the algorithm is a table such that in each row:

$$\text{sign}\left[\text{corr}(a_1, a_2)\right] = \text{sign}\left[\text{corr}(a_1, a_2 | c = s)\right] \quad (1)$$

Simpson's Paradox in Partial Data

We propose that subsampling the data may allow less computationally expensive detection than computing in the whole dataset. We use subsamples sizes of 10%, 30%, 50%, 60%, and 90% of records to assess the accuracy of our approach. For each subsample size, we draw five samples and run our Simpson's paradox detection algorithm. Using the algorithm's result on the whole dataset as the ground truth, we evaluate the performance of the algorithm on the subsets as shown in Figure 2.

The experiment indicates that our algorithm can achieve a high F_1 score in a subset of the data, implying that our method has potential utility in streaming data scenarios.

Algorithm 1 Simpson's Paradox Detection Algorithm

```
INPUT: Relational Table R
con_col ← detectTypes(R)
cat_col ← detectTypes(R)
for all (col1,col2) ∈ con_col do
  corrMatrix1 ← computeCorrelation(col1,col2)
end for
for col ← cat_col do
  subgroups ← R.groupby(col)
  for group ← subgroups do
    for all (col1,col2) ∈ con_col do
      corrMatrix2 ← computeCorrelation(col1,col2)
    end for
    if isReverse(corrMatrix1, corrMatrix2) then
      SP_result ← subgroup_info
    end if
  end for
end for
```

		Attribute 1	*Attribute 2*
Blue	*Attribute 1*	1.0000	−0.6190
	Attribute 2	−0.6190	1.0000
Red	*Attribute 1*	1.0000	−0.6160
	Attribute 2	−0.6160	1.0000

Table 1: Per-group Correlation matrices for the synthetic data set.

Experiments

We perform experiments on a synthetic data set and two data sets from University of California, Irvine machine learning repository (Lichman, 2013): Iris (Fisher, 1936) and Auto Miles per Gallon (Quinlan, 1993).

Synthetic Data Set

As a preliminary validation of our algorithm, we generate 100 records of synthetic data as shown in Figure 1. We manually set means and subgroup-shared covariance matrix for generating samples from a multivariate normal distribution that induces the Simpson's paradox. As shown in Table 2, our detection algorithm finds full Simpson's Paradox for color.

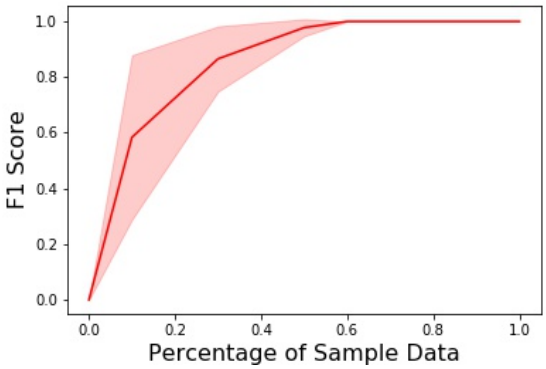


Figure 2: The F_1 score when running the Simpson's paradox over random samples of data.

allCorr	attr1	attr2	revCorr	catAttr	subgroup
0.7710	attribute 1	attribute 2	−0.6190	color	b
0.7710	attribute 1	attribute 2	−0.6160	color	r

Table 2: Result from our algorithm for the synthetic data set. Equation 1 holds on each row

Iris Data Set

The Iris data set has 150 records of 5 attributes: sepal length, sepal width, petal length, petal width, and species. The first four attributes are continuous valued measurements and species is categorical with three values.

Our algorithm detects nine trend reversals, shown in the Table 3. Simpson's Paradox exists with respect to three pairs of measurements (sepal length vs. sepal width, sepal width vs. petal length, and sepal width vs. petal width), since all three species have opposite trends from the population as visualized in Figure 3.

Auto MPG Data Set

From the Auto MPG data set, we select three continuous (mpg, acceleration, and horsepower) and three categorical attributes (cylinders, model year, and origin) and retain only the 392 complete records. As shown in Table 4, we detect six occurrences of Simpson's paradox; four with respect to cylinders and two with respect to model year.

allCorr	attr1	attr2	revCorr	catAttr	subgroup
−0.1090	sepal length	sepal width	0.7470	class	setosa
−0.1090	sepal length	sepal width	0.5260	class	versicolor
−0.1090	sepal length	sepal width	0.4570	class	virginica
−0.4210	sepal width	petal length	0.1770	class	setosa
−0.4210	sepal width	petal length	0.5610	class	versicolor
−0.4210	sepal width	petal length	0.4010	class	virginica
−0.3570	sepal width	petal width	0.2800	class	setosa
−0.3570	sepal width	petal width	0.6640	class	versicolor
−0.3570	sepal width	petal width	0.5380	class	virginica

Table 3: The output from our algorithm for Iris data set. (Equation 1 is true)

allCorr	attr1	attr2	revCorr	catAttr	subgroup
0.4230	mpg	acceleration	−0.8190	cylinders	3
0.4230	mpg	acceleration	−0.3410	cylinders	6
0.4230	mpg	acceleration	−0.0510	model year	75
0.4230	mpg	acceleration	−0.0510	model year	79
−0.7780	mpg	horsepower	0.6210	cylinders	3
−0.7780	mpg	horsepower	0.0130	cylinders	6

Table 4: The output from our algorithm for Auto MPG data set, (Equation 1 is true for each row)

		10 attr.	20 attr.	30 attr.
100K	32 Clu.	4.383	11.499	28.723
	256 Clu.	5.144	14.512	33.954
	1024 Clu.	10.797	24.270	54.154
500K	32 Clu.	5.544	16.033	38.259
	256 Clu.	6.815	18.703	44.084
	1024 Clu.	12.272	29.723	63.196
1M	32 Clu.	6.855	22.303	52.011
	256 Clu.	8.165	23.965	55.423
	1024 Clu.	13.811	34.985	76.289

Table 5: Running time (in s) of detection algorithm

Time Evaluation

We implement our algorithm in Python and run in a Jupyter Notebook on a MacBook Pro with a 2.7 GHz Intel Core i5 processor and 8GB 1867MHz DDR3 RAM to evaluate time.

We keep the two continuous attributes and a categorical attribute that induce Simpson's paradox. In the Table 5, 32 clusters means that there are 32 subgroups partitioned by the categorical attribute. To evaluate performace in varied data sizes, we generate equal numbers of extra continuous attributes (random Gaussian) and categorical attributes (uniformly random integers). We generate synthetic data set three times for each test case and report the average run time.

There are three important factors that influence the run time of our algorithm from our experiments: the total number of attributes, the total number of records, and the number of levels for each categorical attribute.

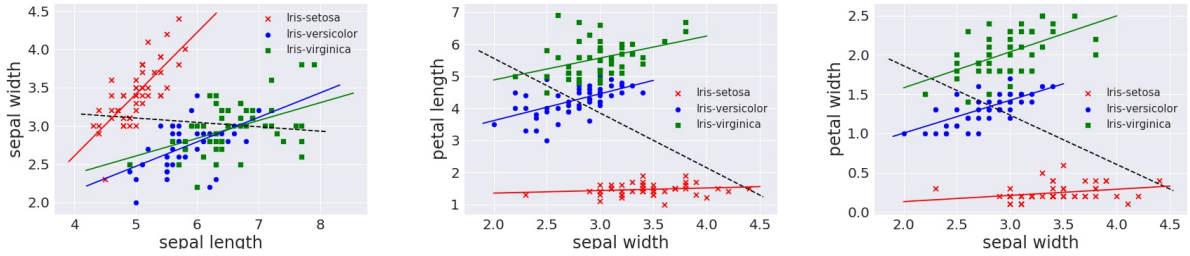
Discussion

In the analysis of real data sets, we found instances of both full (trend reversal for all values of the conditional variable) and partial (reversal for some values) Simpson's Paradox.

We noted that in some detection of Simpson's Paradox, the relationship between two continuous attributes in the whole data set was a strong, while the subgroup relationship was reversed and weak, for example 6 cylinders line in Table 4). This suggests that a distance-based detection may be important in an approximate algorithm.

Conclusions and Future Work

We present a new approach for detecting Simpson's paradox based on the correlation comparison. Our case study on the empirical data sets showed that our algorithm is effective.



(a) sepal length vs. sepal width (b) sepal width vs. petal length (c) sepal width vs. petal width

Figure 3: Simpson's paradox in the Iris dataset. Dashed lines show the overall trends, solid show the individual species.

Further, we explore the feasibility of detecting Simpson's paradox in subsampled data as a preliminary step toward improved scalability. Empirical runtime results confirm that the total number of continuous attributes and categorical attributes, the total number of records, and the levels for each categorical attribute influence the running time of our algorithm from our experiments.

In our current implementation, we partition the data only once and iterate the partition by different categorical attributes on the entire data set. Grouping on two or more columns simultaneously may be necessary to thoroughly detect all surprising results of this form. We want to develop new visual and interactive techniques that use Simpson's paradox to guide a user's data exploration.

Acknowledgements

The FAA has partially sponsored this project through the Center of Excellence for Technical Training and Human Performance. The agency neither endorses or rejects the findings of this research.

References

Alin, A. 2010. Simpson's paradox. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(2):247–250.

Arah, O. A. 2008. The role of causal reasoning in understanding simpson's paradox, lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology* 5(1):5.

Armstrong, Z., and Wattenberg, M. 2014. Visualizing statistical mix effects and simpson's paradox. *IEEE transactions on visualization and computer graphics* 20(12):2132–2141.

Bandyoapdhyay, P. S.; Nelson, D.; Greenwood, M.; Brittan, G.; and Berwald, J. 2011. The logic of simpson's paradox. *Synthese* 181(2):185–208.

Bickel, P. J.; Hammel, E. A.; O'Connell, J. W.; et al. 1975. Sex bias in graduate admissions: Data from berkeley. *Science* 187(4175):398–404.

Blyth, C. R. 1972. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 67(338):364–366.

Chen, A.; Bengtsson, T.; and Ho, T. K. 2009. A regression paradox for linear models: Sufficient conditions and relation to simpson's paradox. *The American Statistician* 63(3):218–225.

Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Schieber, S.; Waldo, J.; Weinberger, D.; and Wood, A. 2017. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Fabris, C. C., and Freitas, A. A. 2000. Discovering surprising patterns by detecting occurrences of simpson's paradox. In *Research and Development in Intelligent Systems XVI*. Springer. 148–160.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of human genetics* 7(2):179–188.

Freitas, A. A. 1998. On objective measures of rule surprisingness. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, 1–9. Springer.

Froelich, W. 2013. Mining association rules from database tables with the instances of simpson's paradox. In *Advances in Databases and Information Systems*, 79–90. Springer.

Guo, Y.; Binnig, C.; and Kraska, T. 2017. What you see is not what you get!: Detecting simpson's paradoxes during data exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 2. ACM.

Hernán, M. A.; Clayton, D.; and Keiding, N. 2011. The simpson's paradox unraveled. *International journal of epidemiology* 40(3):780–785.

Kievit, R. A.; Frankenhuis, W. E.; Waldorp, L. J.; and Borsboom, D. 2013. Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology* 4.

Lerman, K. 2017. Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science* 1–10.

Lichman, M. 2013. UCI machine learning repository.

Pavlides, M. G., and Perlman, M. D. 2009. How likely is simpson's paradox? *The American Statistician* 63(3):226–233.

Pearl, J. 2011. Simpson's paradox: An anatomy. *Department of Statistics, UCLA*.

Quinlan, J. R. 1993. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, 236–243.

Rücker, G., and Schumacher, M. 2008. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC medical research methodology* 8(1):34.