# Redhyte:
# An Interactive Platform for Rapid Exploration of Data and Hypothesis Testing


A thesis submitted by
**TOH WEI ZHONG**
in partial fulfilment for the
Degree of Bachelor of Science with Honours
in
Computational Biology

Supervisors:
Associate Professor **CHOI KWOK PUI**
Professor **WONG LIMSOON**

Semester 1, 2014/2015

## Acknowledgements

I would like to express my gratitude to Associate Professor Choi Kwok Pui and Professor Wong Limsoon for the year-long opportunity to learn from and work under them. Without their guidance and valuable insights, this thesis and the developing of our system would not have been possible. Their combined prowess and experience with statistics and data mining has opened my eyes, as I took away numerous lessons from them throughout the course of the project. It was a well-spent academic year.

## Table of Contents

## Abstract

Hypothesis testing is a well-developed framework in data analysis. Conventionally, data is collected with a scientific question in mind, from which a formulated hypothesis is tested using statistical tests. In the current "Big Data" context, data is more often collected and stored without any scientific question a priori. Furthermore, having large number of attributes in a dataset renders hypothesis testing to be practically lacklustre and methodologically unsound, as it is not possible for an analyst to purposefully examine and scrutinise the entire dataset. Motivated as such, we developed Redhyte, an interactive platform for rapid exploration of data and hypothesis testing. We first assess the adequacy of statistical tests that the user is interested in. Next, we augment data mining techniques, specifically supervised and class-imbalance learning, to the hypothesis testing framework, in a bid for more wholesome and efficient hypothesis testing. We termed the chief objectives of Redhyte "hypothesis analysis" and "hypothesis mining" – the search for interesting hypotheses in a dataset. Hypothesis mining consists of three steps: "context mining", mined hypothesis formulation, and mined hypothesis scoring on interestingness. To capture and evaluate specific aspects of interestingness, we developed and implemented various hypothesis mining metrics. Finally, we give an illustration how Redhyte can be used to enrich the arsenal of the scientist and the data analyst. Redhyte is an R shiny web application and can be found online at https://tohweizhong.shinyapps.io/redhyte/, with the source codes housed in a GitHub repository at https://github.com/tohweizhong/redhyte.



Keywords: statistical hypothesis testing, hypothesis analysis, data mining, random forest, hypothesis mining

# Introduction

## 1.1 Statistical hypothesis testing

Data analysis has long been an integral part of any serious pursuit – from science to business, it is impossible to isolate oneself from the science and art of data analysis. Fundamentally, data analysis allows us to i) validate an idea or a hunch, ii) uncover new information and knowledge from data, and iii) disprove ostensible phenomena, all of which encapsulated in the formal framework of data analysis.

Despite the mathematical and statistical underpinnings of data analysis, it is not wrong to claim that data analysis is in fact an art (Agresti and Franklin, 2012; Jarman, 2013). It is one thing to be learning mathematical statistics and data mining – understanding the asymptotic behaviour of parameter estimates, learning the formal and intricate details of statistical tests, and comprehending the subtleties of classification models, it is yet another thing to be able to look at empirical and raw data and from it, extract new information and knowledge. The art of data analysis has long been a domain-driven endeavour that combines statistical proficiency with domain-specific experience, intuition and subject matter, hence an art.

One of the data analyst's primary tools is that of comparison, or more formally known as hypothesis testing (Froehlich and Kent, 1995; Engle, 1984). Putting together a hypothesis with a statistical test allows the data analyst to make justifiable conclusions from the data, based on the results of the test. Moreover, the collection of data could be driven by the initial question or hypothesis in mind. Such a conventional, domain knowledge- and hypothesis-driven approach has served us well thus far.

## 1.2 Data analysis in Big Data context

However, with the advent of Big Data, analysing data in such a conventional manner may not be feasible or even possible. In a typical setting before Big Data, the data analyst starts with a domain knowledge-driven question in mind, collect the relevant data if it is not already available, formulate the hypothesis, test the hypothesis, and make a conclusion. The Big Data circumstance that we are currently in brings about two interesting scenarios, specifically the collection of data without a scientific question a priori, and the "large *p* small *n*" phenomenon (West, 2003). With centralized storage, high-quality curation, and convenient retrieval and dissemination of data, data has been routinely collected without a given set of questions in mind. In the biomedical and healthcare context for instance, modern high-throughput assaying technologies and better storage and curation of electronic medical and healthcare records are giving us more, cleaner data that may not be collected with a prior hypothesis in mind. These data contains plenty of useful knowledge, waiting to be unearthed. This is likewise for areas outside of the sciences, such as business and the social sciences.

Moreover, having a large number of attributes in a dataset requires adequate treatment and analysis to properly account for these attributes. Consider this: formulating a hypothesis concerning a small number of attributes and testing it in a large dataset while ignoring the other attributes is not only wasteful but flawed (due to issues such as violation of statistical assumptions and confounding). For example, given a hypothesis concerning two attributes, say *A* and *B*, for a certain class of a third categorical attribute *C*, the initial hypothesis could be amplified, i.e. the trend observed between *A* and *B* is strengthened when we consider the certain class of *C*. The trend could also be reversed; this is commonly known as Simpson's Reversal (Pavlides and Perlman, 2009). A conventional, domain knowledge-driven approach of analysing data gives no simple or systematic way to reveal such phenomena, leaving

discoveries of such to intuition and chance. An epitome of such a phenomenon is the UC Berkeley gender bias case (Bickel *et al*, 1975).

### 1.3 Data mining

Data mining is a well-established class of techniques commonly used to search for interesting and global relationships in the large datasets. For instance, supervised classification models such as decision trees allows for the classification of a target or response attribute based on other attributes (Mylers *et al*, 2004), while unsupervised clustering models such as k-means clustering construct empirical and observed "cliques" or clusters in the data (Steinley, 2006). Frequent pattern mining techniques mine for recurring patterns in the data, and thus reveal any form of associations present (Goethals, 2003). While these techniques mine for relationships between attributes based on how they confer to the classification or clustering of data, they do not contribute directly to the fundamental endeavour of making comparisons – knowing that an certain attribute *A* contributes greatly to the classification of a response attribute of interest *B* is not nearly as intuitive as putting both *A* and *B* in a contingency table, as in Table 1. This is especially so for those not trained in statistics or data mining. A concrete example would be in genomics, where microarray data can be used to identify the genes that, if up- or down-regulated, contribute to the classification of phenotype. After using a classification model to identify these genes, it is often apt to make use of hypothesis testing as downstream analysis step to understand exactly how these genes contribute to the observed phenotype (Smyth *et al*, 2003).

Table 1: An example contingency table

| **Gene A** | **Diseased** | **Control** | **Total** |
|---|---|---|---|
| **Up-regulated** | 43 | 27 | 70 |
| **Not up-regulated** | 12 | 44 | 56 |
| **Total** | 55 | 71 | 126 |

## 1.4 Motivation and results

Therefore, in this piece of work we have developed a system named Redhyte, an interactive platform for "Rapid Exploration of Data and Hypotheses Testing". Redhyte stands in the middle ground between domain knowledge-driven hypothesis testing and data-driven data mining on entire datasets: based on a rough domain knowledge-driven hypothesis that the user has in mind, Redhyte searches for related hypotheses that could enhance or negate the initial hypothesis (section 2.1.5). Redhyte utilizes data mining techniques for such a search. Moreover, Redhyte is able to assess the adequacy of statistical tests, and analyse hypotheses in order to pinpoint main significance contributors (section 2.1.4). Redhyte was designed and developed for datasets with large numbers of attributes, possibly collected without any scientific questions a priori. Redhyte requires the user to conjure an initial hypothesis consisting of a small number of attributes, based on intuition or experience, in a large and possibly unexplored dataset. Using this user-defined initial hypothesis, Redhyte generates hypotheses that are potentially interesting to the user.

Redhyte consists of a core algorithm and a graphical user interface (GUI) for the user to utilise the algorithm. The primary objectives of Redhyte are to conduct "hypothesis analysis" and "hypothesis mining" – to mine for valid and interesting hypotheses based on the user's initial hypothesis. Whether a hypothesis is interesting and hence sieved out is defined in Redhyte's algorithm, and will be elaborated in a later section (section 4.5.1).

## 1.5 Related work

Fundamentally, Redhyte serves a tool to enhance data analysis, or more specifically, hypothesis testing and exploratory data analysis. Consider the typical analysis workflow of a data analyst: given a dataset, after some descriptive statistics, the analyst may start off with

the testing of some simple hypotheses that are naturally intuitive. Redhyte contributes to this endeavour by not only allowing the analyst to conduct such exploratory hypothesis testing, Redhyte also allows the analyst to seamlessly transit to hypothesis mining, based on the initial exploratory hypotheses.

Moreover, while statistical software such as R (R Core Team, 2014) allow the user to easily utilise hypothesis testing, they are unable to do automated checks of whether the tests are applied correctly – if the assumptions of the test, such as equal variances or identical distributions (i.i.d.), are not met by the data, the test and any conclusions drawn from it are no longer valid. This is especially pertinent in the current Big Data setting, as opposed to a more conventional setting of a scientific study. For example, in a traditional cohort or cross-sectional study, subjects are carefully selected such that assumptions of any statistical tests or models that are to be used are met in the collected data. On the other hand, if data is routinely collected without any major scientific question a priori, then it is easy for statistical assumptions to be dissatisfied. In Redhyte, such checking of assumptions is automated. We make a more in-depth comparison of Redhyte with other software systems in section 3.1.1.

Hypothesis mining without the consideration of a user's initial hypothesis has been done by Liu *et al* (2011). Liu *et al*'s hypothesis mining system employs frequent pattern mining techniques to search for significant and interesting hypotheses, by representing a hypothesis' subpopulations as patterns in the data. In comparison to Redhyte, Liu *et al*'s system is entirely data-driven as the latter does not interact with the user's domain expertise. There are pros and cons to such an approach, to be discussed in a later section (section 4.1).

## Materials & Methods

Prototyping of the system is done using the R programming language and the *shiny* package (Chang, 2015). In order to validate Redhyte's algorithm and assess its performance, we used well-known datasets from the UCI Machine Learning Repository (Lichman, 2013), namely *adult*, *mushroom*, and *arrhythmia*, and the UC Berkeley admission dataset (Bickel *et al*, 1975).

### 2.1 Framework

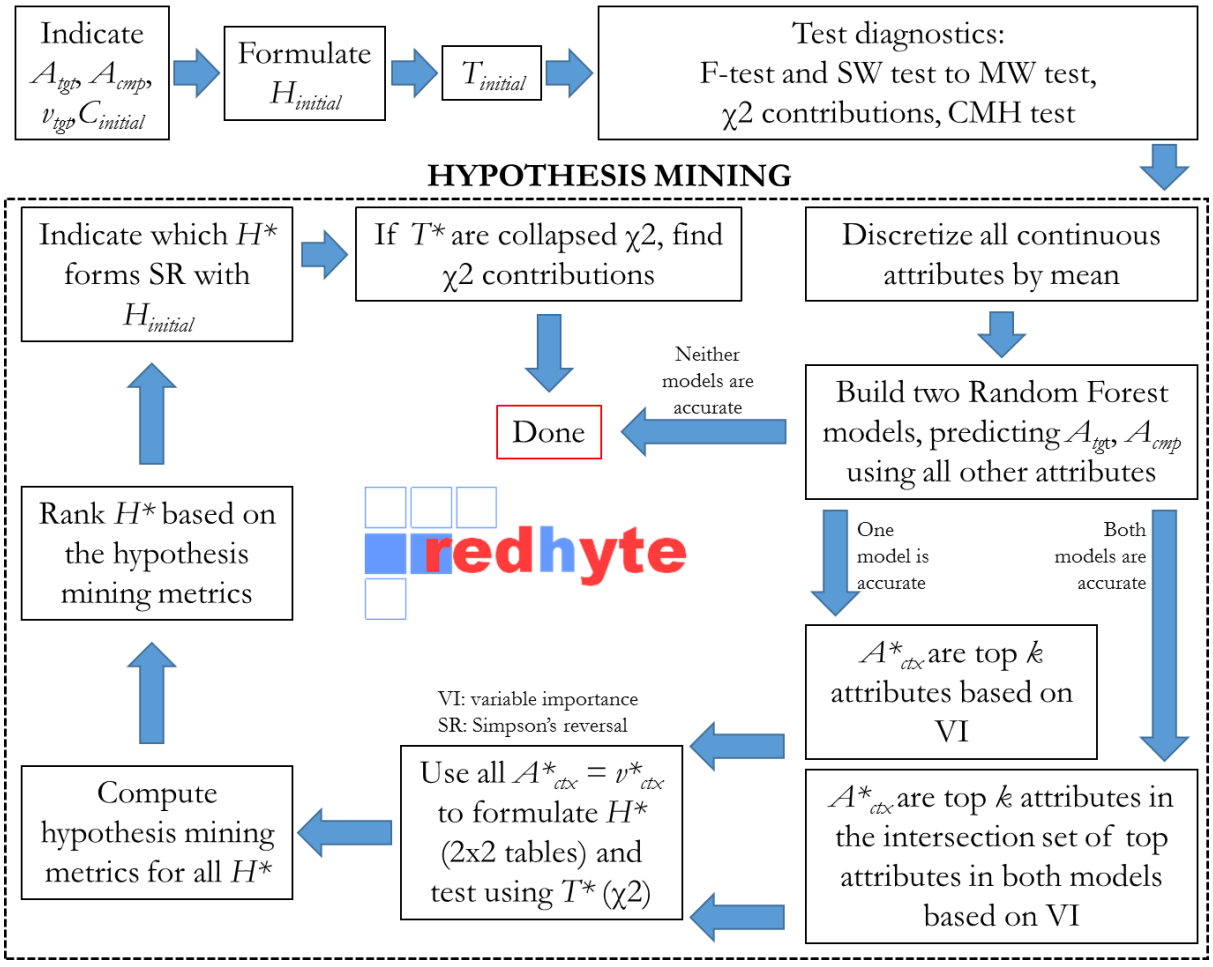



Figure 2: Overview of the inner workings of Redhyte

Figure 2 depicts a high-level view of how Redhyte conducts hypothesis mining. The shorthand notations used in the figure are as follows: $A_{tgt}$ for target attribute, $A_{cmp}$ for comparing attribute, $v_{tgt}$ for target attribute value, $C_{initial}$ for initial context, $H$ for hypothesis,

*T* for test, and finally, the SW, MW, and CMH test for the Shapiro-Wilk, Mann-Whitney, and the Cochran-Mantel-Haenszel test respectively. These terms and statistical tests will be further expounded later in the section. We first describe the well-understood framework of hypothesis testing, and later augment it with Redhyte's hypothesis mining framework.

### 2.1.1 Hypothesis testing

A statistical hypothesis, in lay terms, is a comparison of difference between *n* subsets of a dataset, where *n* is positive. We call these subsets *subpopulations* of a hypothesis. If the differences between the subpopulations are likely to be real and did not occur by chance, the difference is said to be statistically significant. Statistical significance is attained by subjecting the hypothesis to a statistical test, such as the t-test or $\chi 2$ test.

For $n = 1$, the hypothesis concerns the comparison between a subpopulation of the data and the actual population from which the data is sampled from. In reality such true population parameters are often unknown. On the other hand, for $n > 2$, the comparison is across multiple subpopulations. It is often hard to utilise and interpret the results of such a comparison. Ultimately, users may still need to consider pairwise comparisons. Due to these reasons, Redhyte only considers hypotheses where $n = 2$.

Typically, the process of hypothesis testing consists of several steps:

1. Based on available data, state the null and alternative hypothesis, $H_0$ and $H_1$. $H_0$ typically asserts that there is no difference in the subpopulations, while $H_1$ asserts its negation.
2. Based on the type and distribution of the attribute that the subpopulations are being compared on, select an appropriate test. For instance, if the subpopulations are being

compared on body mass index (BMI) and the distributions of BMI is approximately Normal, the Student's t-test can be used. On the other hand, if the subpopulations are being compared on a dichotomous attribute of overweight versus not overweight, a $\chi 2$ test would be appropriate.

3. Using the selected test, compute a p-value. The p-value is the probability of observing the sample data, or a more extreme sample, given that the $H_0$ is true. In other words, if the probability of observing the collected data is very low given that $H_0$ is true, then there is some evidence to reject $H_0$ and hence accept $H_1$. Typically the decision to accept or reject $H_0$ is dependent on an arbitrary threshold known as the significance level, which is usually 0.05 or 0.01.
4. Accept or reject $H_0$ based on the computed p-value and the significance level.

### 2.1.2 Initial hypothesis

The first step in Redhyte's hypothesis mining system is for the user to set up an initial hypothesis. In this section we describe the terminologies in hypothesis testing, as used in Redhyte. Several of the following terms come from Liu *et al*, with an extension of terminologies unique to Redhyte, to be introduced later. We use the following toy dataset as a running example to illustrate the terminologies in Redhyte:

Table 3: Toy dataset to illustrate Redhyte's hypothesis mining framework

| ***ID*** | ***Gender*** | ***Native country*** | ***Income*** | ***Family history for cardiac disease*** | ***Smoking status*** | ***Resting Heart Rate (numerical)*** | ***Resting Heart rate (categorical)*** |
|---|---|---|---|---|---|---|---|
| **1** | M | S'pore | Low | False | Smoker | 75 | High |
| **2** | F | S'pore | Low | False | Never | 65 | Low |
| **3** | F | M'sia | High | False | Smoker | 81 | High |
| **4** | M | S'pore | High | True | Quitter | 72 | High |
| **…** | … | … | … | … | … | … | … |
| **N** | M | China | High | True | Quitter | 73 | High |

Consider the following hypotheses:

Table 4: Example hypotheses

| | |
|---|---|
| Hypothesis A | Do smokers have a higher resting heart rate than the never-smokers and quitters? |
| Hypothesis B | Out of all the men, do smokers have a higher resting heart rate than the never-smokers and quitters? |

In a hypothesis, we define the *target attribute* as the attribute that represents the result, response, or outcome. In both hypotheses A and B, the target attribute is resting heart rate. The target attribute may be numerical or categorical, of which influences the type of statistical tests used to assess the hypothesis. In our examples, resting heart rate may be numerical (e.g. 75 beats per minute) or categorical (e.g. high versus low). Furthermore, for a categorical target attribute, we define a *target attribute value* as a group within the categorical target attribute that is the most meaningful or interesting; this often pertains to an affirmative (e.g. diseased being more interesting than not diseased, exposed being more interesting than control), though it is context-dependent. On the other hand, the *comparing attribute* is the attribute that represents the act of comparison and/or intervention. In our examples, the comparing attribute is smoking status. Comparing attributes must be categorical.

Hypothesis B considers a smaller subpopulation, by restricting the samples of the dataset to those that satisfies the condition of gender being male. In Hypothesis B, we call gender a *context attribute*, and the condition of {Gender = M} as a *context item*. In general, more context items, such as {Family history for cardiac disease = True} or {Native country = S'pore}, can be added into the hypothesis. The set of context items that are added into the initial hypothesis is known as the *initial context*. The more context attributes there are in the initial context, the more specific the hypothesis becomes, and the smaller its subpopulations.

### 2.1.3 Statistical tests and contingency tables

After the initial hypothesis is set up, a statistical test is used to assess the hypothesis. Choosing the correct test is largely dependent on the type and distribution of the target attribute. Motulsky (2014) gave a concise overview of the different variants of statistical tests. Of the different statistical tests, Redhyte only uses the t-test (Student, 1908) and the $\chi 2$ test (Pearson, 1900) to assess the initial hypothesis. The exact variants of t-test and $\chi 2$ test used in Redhyte are the Welch 2-sample t-test (Welch, 1974), which assumes unequal variances for the subpopulations, and the Pearson's $\chi 2$ test with Yate's continuity correction (Yates, 1934). The Yate's continuity correction is often used in the computation of the $\chi 2$ test statistic to compensate for the fact that the $\chi 2$ distribution, being a continuous distribution, is used to approximate probabilities in discrete data.

A simple way to represent hypothesis is to use tables. To illustrate, Hypothesis A can be represented in the following *contingency table*:

Table 5: Contingency table of Hypothesis A

| | **High resting heart rate** | **Low resting heart rate** | **Total** |
|---|---|---|---|
| **Smokers** | $c_{11}$ = 43 (61.4%) | $c_{12}$ = 27 (38.6%) | 70 |
| **Never-smokers and quitters** | $c_{21}$ = 12 (21.4%) | $c_{22}$ = 44 (78.6%) | 56 |
| **Total** | 55 | 71 | 126 |

In this hypothetical example, each $c_{ij}$ represents the number of samples in the data that are in the $i^{th}$ and $j^{th}$ group. In the contingency table, the columns are used to represent the target attribute, while rows for the comparing attribute. The percentages in the each cell of the table represent the proportions of samples in a group of the comparing attribute that belongs to a group of the target attribute. Note that a target or comparing attribute group may consist of multiple classes from that attribute. For instance, the never-smokers and quitters form a single group in the hypothesis.

Alternatively, Hypothesis A can also be represented in a table to compare means:

Table 6: Comparison table of Hypothesis A

| | Means of resting heart rate | Standard deviation of resting heart rate | Support | Proportions |
|---|---|---|---|---|
| **Smokers** | 75 | 20 | 70 | 56% |
| **Never-smokers and quitters** | 65 | 18 | 56 | 44% |

We call the above table a *comparison table*, where the means of the target attribute are being compared across the groups of the comparing attribute. The supports for each group of the comparing attribute are simply the number of samples that belong to the respective groups. For a given hypothesis, depending on how the target attribute is being represented in the dataset (numerical or categorical), Redhyte generates the appropriate table.

In addition, Redhyte discretizes numerical target attributes by its mean, to give a binary target attribute of above or below mean. This binary target attribute will be used later in hypothesis mining, as well as to generate a contingency table for the initial hypothesis. Therefore, for both numerical and categorical target attributes, contingency tables are generated, in addition to the comparison table for numerical target attributes. Redhyte proceeds subsequently to perform the initial test(s) on the given hypothesis, which are the t-test and/or the $\chi 2$ test.

### 2.1.4 Test diagnostics and hypothesis analysis

After the initial hypothesis and test is set up, Redhyte proceeds to do test diagnostics and hypothesis analysis. For numerical target attributes, the Student's t-test, or simply t-test, is used to assess the initial hypothesis. The t-test is a parametric test that requires certain assumptions to hold true, such as normality and equal variances (Havlicek and Peterson, 1974). In order to assess whether these assumptions are met, Redhyte uses the Shapiro-Wilk test (Shapiro and Wilk, 1965) and the F-test (Box, 1953). For the Shapiro-Wilk test, the null

hypothesis states that the target attribute is normally distributed in both subpopulations of the initial hypothesis, while that of the F-test states that the two subpopulations under comparison have the same variance. If any of these tests are significant, Redhyte uses the non-parametric Mann-Whitney test (Mann and Whitney, 1947), also known as the Wilcoxon rank sum test, to assess the initial hypothesis. Both the normality and equal variances assumptions form part of the general independent and identically distributed (i.i.d.) assumption, an assumption found in many, if not all, statistical tests. We use the Cochran-Mantel-Haenszel test (to be elaborated later in the section) in an attempt to address the independence assumption, though this assumption is generally not assessable.

For categorical target attributes, the $\chi 2$ test was used to assess the initial hypothesis. Unlike the t-test, the $\chi 2$ test is non-parametric, with a null hypothesis stating that there is no association found between the categorical target and comparing attributes in the contingency table. In Redhyte, we define a *collapsed $\chi 2$ test* as a $\chi 2$ test whereby one or both of the groups of the comparing attribute comprises of more than one comparing attribute class – a $\chi 2$ test on Table 5 would be a collapsed $\chi 2$ test. In contrast, a $\chi 2$ test on the following contingency table would be a *flat $\chi 2$ test*:

Table 7: Flat contingency table of Hypothesis A

| | **High resting heart rate** | **Low resting heart rate** | **Total** |
|---|---|---|---|
| **Smokers** | $c_{11}$ = 43 (61.4%) | $c_{12}$ = 27 (38.6%) | 70 |
| **Never-smokers** | $c_{21}$ = 6 (20.7%) | $c_{23}$ = 23 (79.3%) | 29 |
| **Quitters** | $c_{31}$ = 6 (22.2%) | $c_{32}$ = 21 (77.8%) | 27 |
| **Total** | 55 | 71 | 126 |

The comparing attribute group of the never-smokers and quitters are separated into their individual classes in a *flat contingency table*. The rationale behind a flat $\chi 2$ test in hypothesis analysis is to find the *$\chi 2$ top contributor* – the class within the comparing attribute that

contributes most to the $\chi 2$ test statistic, and hence the test's significance. The $\chi 2$ test statistic is given by

$$\chi 2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where $O_{ij} = c_{ij}\ \forall i, j$ and $E_{ij}$ is the expected or theoretical count, asserted by the null hypothesis of independence. For example, the $\chi 2$ test statistic for Table 7 is 20.25 with $p < 0.05$, varyingly contributed by classes of the comparing attribute as follows:

Table 8: Computation for $\chi 2$ contributions in Hypothesis A

| **High resting heart rate** | **Observed** | **Expected** | $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ | **$\chi 2$ contribution** |
|---|---|---|---|---|
| **Smokers** | $O_{11} = 43$ | $E_{11} = 30.6$ | 5.02 | 24.8% |
| **Never-smokers** | $O_{12} = 6$ | $E_{12} = 12.7$ | 3.53 | 17.5% |
| **Quitters** | $O_{13} = 6$ | $E_{13} = 11.8$ | 2.85 | 14.1% |

The $\chi 2$ contributions are only evaluated for the target attribute value, which in this case is high resting heart rate. Here, the top contributor is the smoker class, i.e. the smokers are the main reason why the $\chi 2$ test on Table 7 is significant.

To further motivate the identification of the $\chi 2$ top contributor, consider the following flat contingency table on various vaccines (example courtesy of Wong, 2014):

Table 9: Flat contingency table on vaccine effectiveness

| **Vaccine** | **Type** | **Had flu** | **Avoided flu** | **Total** |
|---|---|---|---|---|
| **A** | Attenuated | 43 (15.4%) | 237 (84.6%) | 280 |
| **B** | Attenuated | 52 (20.8%) | 198 (79.2%) | 250 |
| **C** | Attenuated | 25 (9.3%) | 245 (90.7%) | 270 |
| **D** | Protein subunit | 48 (18.5%) | 212 (81.5%) | 260 |
| **E** | Protein subunit | 57 (19.7%) | 233 (80.3%) | 290 |
| **Total** | | 225 | 1125 | 1350 |

In the example above, a flat $\chi 2$ test would return a p-value of less than 0.05, thereby suggesting that the vaccines are effective against flu incidence. Searching for the top $\chi 2$

contributor amongst the five vaccines (using "Avoided flu" as the target attribute value), we have the following Table 10.

Table 10: Computation for $\chi 2$ contributors of Table 9

| Avoided flu | Observed | Expected | $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ | Contribution |
|---|---|---|---|---|
| **A** | 237 | 233.3 | 0.059 | 0.36% |
| **B** | 198 | 208.3 | 0.509 | 3.07% |
| **C** | 245 | 225 | 1.778 | 10.74% |
| **D** | 212 | 216.7 | 0.102 | 0.62% |
| **E** | 233 | 241.7 | 0.313 | 1.89% |

Vaccine C contributes 10.74% of the $\chi 2$ test statistic, possibly suggesting that vaccine C could be the main reason why the flat $\chi 2$ test is significant. More importantly, it is possible that vaccine C is the most effective vaccine of the five. Indeed, we can make the following comparison, and with the $\chi 2$ test on Table 11 being significant, it is suggested that vaccine C is the most effective vaccine of all.

Table 11: Contingency table comparing vaccine C with the others

| Vaccine | Had flu | Avoided flu | Total |
|---|---|---|---|
| **C** | 25 (9.3%) | 245 (90.7%) | 270 |
| **A, B, D, E** | 200 (18.5%) | 880 (81.5%) | 1080 |
| **Total** | 225 | 1125 | 1350 |

Finally, Redhyte uses the Cochran–Mantel–Haenszel (CMH) test to point out attributes in the dataset that potentially confounds the initial hypothesis (Mantel, 1963). Given the binary target and comparing attributes and a third stratifying categorical attribute in the data, the CMH test first constructs *k* 2x2 contingency tables, where *k* is the number of classes or strata in the third attribute. Using the *k* 2x2 tables, CMH test asserts in its null hypothesis that the target and comparing attribute are conditionally independent, i.e. in all of the 2x2 tables, the target and comparing attribute are independent of each other. Therefore, for a given third attribute in the dataset, if the null hypothesis is rejected, then the third attribute is said to be a potential confounder and warrants investigation. In addition, we may say that the

independence assumption in the initial test is violated. Redhyte generates a list of such attributes only for the user's consideration – these attributes are not involved in hypothesis mining, the next step in Redhyte's workflow.

### 2.1.5 Hypothesis mining

In the preceding sections, Redhyte is engaging the user with only the initial hypothesis. The next procedure in the Redhyte is known as *hypothesis mining*, and this is where Redhyte takes the entire dataset into consideration. Hypothesis mining consists of three steps: *context mining*, mined hypothesis formulation, and mined hypothesis scoring and ranking.

#### 2.1.5.1 Context mining

Context mining is concerned with the search for attributes in the dataset that may be included in the initial hypothesis as context items. The intuition of context mining is as follows: the initial user-input hypothesis is considered to be domain knowledge-driven, intuitive, and general. We discussed how having additional context attributes in a hypothesis shrinks the subpopulations in a hypothesis, and renders the hypothesis more specific.

The primary objective of Redhyte and hypothesis mining is to be able to automate the search for valid and interesting hypotheses, based on the initial hypothesis. This translates directly to the search for interesting context items in the dataset that can be added into the initial hypothesis. In order to search for such context items, Redhyte first uses classification models to search for attributes in the dataset that contributes to the classification of the target and comparing attribute – if a given attribute, say income, contributes to the classification of the target attribute groups, say high or low resting heart rate, then we say that income is somehow associated with the target attribute. Specifically, adding a particular class in the

income attribute, say {Income = high}, may result in a hypothesis in which either the observed trend in the initial hypothesis is amplified or reversed (Simpson's Reversal). Moreover, for a given categorical attribute *A*, if all classes of *A* form Simpson's Reversals with the initial hypothesis, we say that *A* forms Simpson's Paradox with the initial hypothesis.

In this example we call income and all other attributes sieved out by the classification models as *mined context attributes* (as opposed to the context attributes in the initial context), and these attributes are shortlisted based on variable importance measures in the model. These variables importance measures will be discussed in a later section (section 4.4.2). We note that at this point, it is not possible to tell, for instance, that {Income = high} reverses trend observed in the initial hypothesis, or {Income = low} amplifies the said trend. Prior to context mining, Redhyte discretizes all numerical attributes, including the target attribute if applicable, by the mean. Context mining will be discussed in greater detail in a later section (section 4.4).

#### 2.1.5.2 Formulation, scoring and ranking of mined hypotheses

Using the mined context attributes, Redhyte considers all possible classes within these mined context attributes, and use them as context items to add into the initial hypothesis to form *mined hypotheses*. Redhyte then uses the $\chi 2$ test to evaluate the mined hypotheses. In principle, the information of whether the trend observed in the initial hypothesis has been amplified or weakened in the mined hypotheses is contained in the differences in p-values between the initial test and the test on the mined hypotheses. However, the p-values merely convey the statistical significance of each test – the trend could be reversed from the initial to the mined hypothesis, while retaining the same amount of statistical significance. p-values

alone are unable to sieve out such information, and thus Redhyte relies on four additional *hypothesis mining metrics*, in addition to $\chi 2$ test statistics and p-values, to evaluate the mined hypotheses. These metrics will be discussed in depth in a later section (section 4.5.1).

Since there are multiple hypotheses to be tested at once after the mined hypotheses are formulated, a problem known as multiple testing arises. Multiple testing occurs when there are multiple hypotheses to be tested at one instance, using multiple statistical tests and hence generating multiple p-values. This is leads to an increase of the probability of making one or more false discoveries among the numerous hypotheses, otherwise known as the family-wise error rate (FWER). In Redhyte, the p-values are corrected for multiple testing using the Bonferroni correction (Bonferroni, 1936). The Bonferroni correction is a commonly used technique to correct for FWER, and is as follows: if there are *n* different hypotheses to be tested at one instance, then the FWER can be suppressed by testing each individual hypothesis at a statistical level $\frac{1}{n}$ of what it would be if only one hypothesis was to be tested. Therefore, in all, the mined hypotheses are assessed based on $\chi 2$ test statistics, p-values, adjusted p-values, and four hypothesis mining metrics, a total of seven different metrics.

## Results

### 3.1 Overall design and functionalities

Table 12: A comparison of functionalities Redhyte offers, with other data analysis systems

| Functionalities | Redhyte | Liu *et al* | R/SAS/etc |
|---|---|---|---|
| **User-friendly interface** | ✓ | | |
| **Simple data visualizations** | ✓ | | # |
| **Testing of initial hypothesis** | | | |
| - t-test | ✓ | | # |
| - $\chi 2$ test | ✓ | | # |
| **Test diagnostics and hypothesis analysis** | | | |
| - t-test: automated assumption checks | ✓ | | # |
| - $\chi 2$ test: automated computing of $\chi 2$ contributions | ✓ | | * |
| - Automated CMH test | ✓ | | # |
| **Hypothesis mining** | | | |
| - Automated discovery of Simpson's Paradoxes, etc. | ✓ | ✓ | *^ |
| - Hypothesis mining metrics | 4 metrics | 2 metrics | * |
| **Session log documentation** | ✓ | | |

#: requires programming/scripting
*: novel, not considered by users of these systems
^: difficult, no simple or systematic way to do so in conventional data analysis

Table 12 illustrates a qualitative comparison between Redhyte, Liu *et al*'s hypothesis mining system, and various conventional data analysis systems, with regards to hypothesis testing. One key advantage that Redhyte confers to the endeavour of data analysis is the automation of test diagnostics and hypothesis analysis. For instance, the checking of the normality assumption in the t-test or the identification of the $\chi 2$ top contributor in the $\chi 2$ test requires careful scripting from a data analyst in R. In Redhyte, these diagnostics and analyses are generated automatically.

Furthermore, we introduced novel concepts in Redhyte, such as the notion of a $\chi 2$ top contributor, and the use of classification models in the hypothesis testing framework. $\chi 2$ contributions are generally not considered by users of conventional data analysis systems, unless the user has some intuition on the subject matter, while the use of classification models

facilitates the quick discovery of Simpson's Paradoxes. Finally, the "point-and-click" nature of user interfaces often invokes criticisms of lack of reproducibility. To address this, Redhyte documents and profiles analysis sessions in the session log documentation. The session log can be saved and shared amongst collaborators for reproducibility of results.

### 3.2 User interface

Redhyte is a fundamentally a web application that renders in a web browser, such as Google Chrome or Mozilla Firefox. Redhyte's user-facing interface is organized into tabs, with each tab housing a specific functionality that Redhyte provides. These tabs contain the settings control, data preview, data visualization, initial test module, test diagnostics module, context mining module, mined hypothesis formulation and scoring module, and finally, log documentation.

#### 3.2.1 Settings and Data preview

The first two tabs in the interface are the Settings and the Data Preview tabs. In the Settings tab, users can have specific control over how Redhyte treats the input data. Options such as file types and transposing allow certain degree of flexibility in the data format. Also housed in the Settings tab are settings used in test diagnostics, context mining and hypothesis mining, to give the user more control over the hypothesis mining process. These settings include the maximum number of classes in the categorical attributes of the dataset, the p-value threshold for switching to non-parametric tests in the Test diagnostics module, minimum classification accuracy for context mining models, number of context attributes to mine for, and class-ratio threshold for class-imbalance learning in context mining (refer to section 4.4.3), and minimum cell support for mined hypotheses. The Data preview serves as a simple functionality for users to have a quick peek at the input dataset.

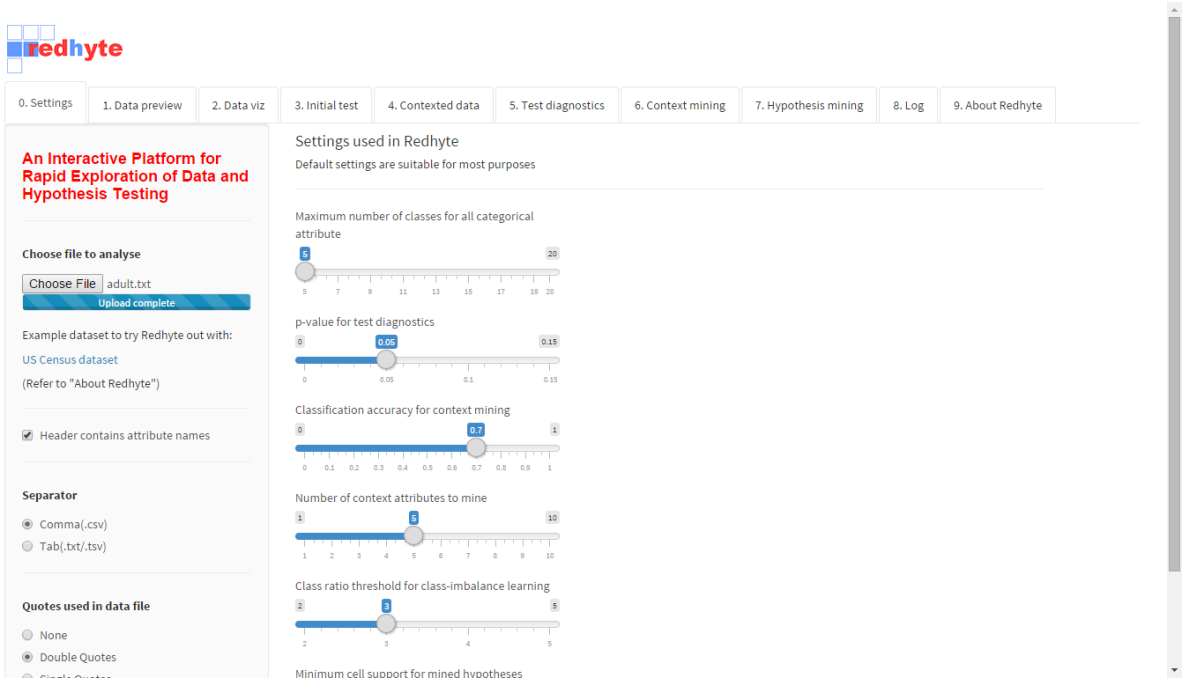
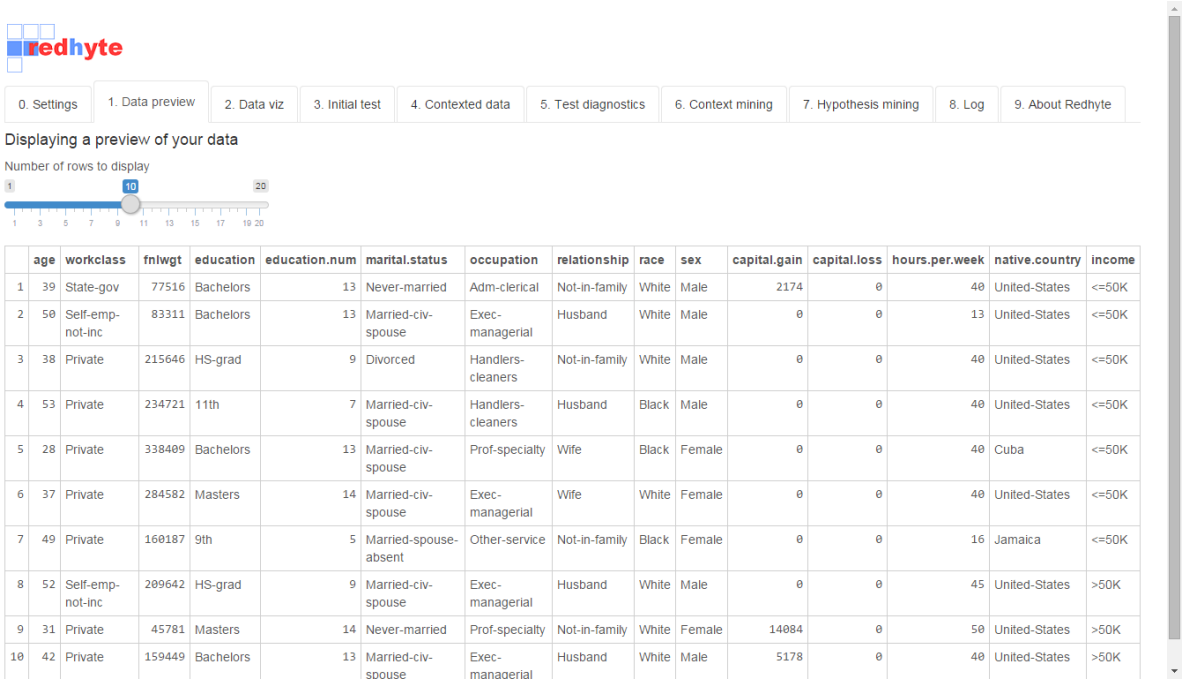



Figure 13: Screenshot of the Settings tab in Redhyte




| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country | income |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 | 40 | United-States | <=50K |
| 2 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 | 13 | United-States | <=50K |
| 3 | 38 | Private | 215646 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 | 40 | United-States | <=50K |
| 4 | 53 | Private | 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0 | 0 | 40 | United-States | <=50K |
| 5 | 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0 | 0 | 40 | Cuba | <=50K |
| 6 | 37 | Private | 284582 | Masters | 14 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 0 | 0 | 40 | United-States | <=50K |
| 7 | 49 | Private | 160187 | 9th | 5 | Married-spouse-absent | Other-service | Not-in-family | Black | Female | 0 | 0 | 16 | Jamaica | <=50K |
| 8 | 52 | Self-emp-not-inc | 209642 | HS-grad | 9 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 | 45 | United-States | >50K |
| 9 | 31 | Private | 45781 | Masters | 14 | Never-married | Prof-specialty | Not-in-family | White | Female | 14084 | 0 | 50 | United-States | >50K |
| 10 | 42 | Private | 159449 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 5178 | 0 | 40 | United-States | >50K |

Figure 14: Screenshot of the Data preview tab in Redhyte

### 3.2.2 Data visualization

The next tab in the interface houses the Data visualization tab. Here, users can select two attributes from the input data, using which Redhyte renders the appropriate statistical graphics for visualisations, such as histograms, barplots, scatterplots, boxplots, and spineplots. The type of statistical graphic rendered depends solely on the type of selected attributes. For example, if the selected attributes are both numerical, a scatterplot is rendered. If the selected attributes are each numerical and categorical, boxplots are rendered, as in Figure 15.

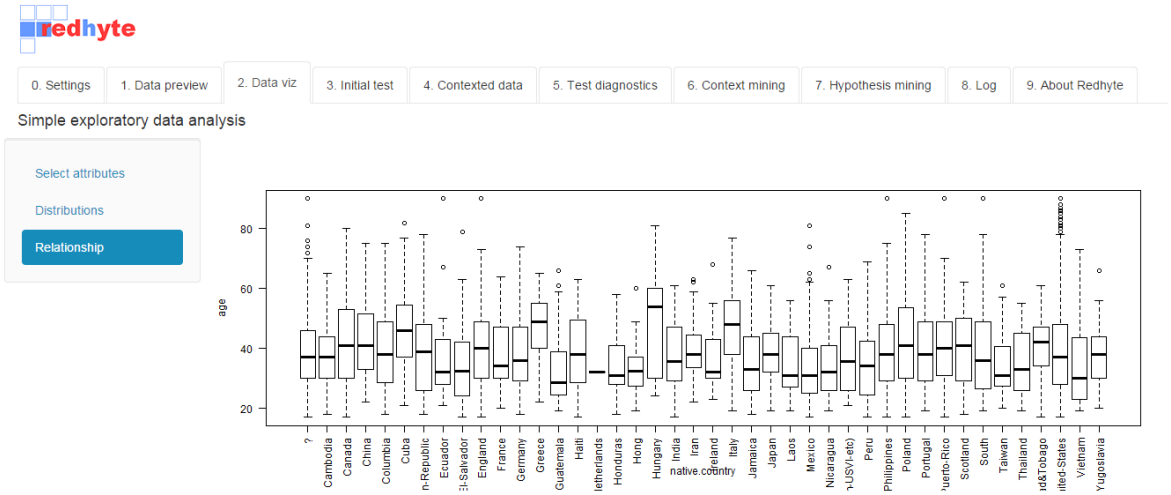



Figure 15: Screenshot of the Data visualization tab in Redhyte

### 3.2.3 Initial test and Test diagnostics, Contexted data

Following the Data visualization tab is the Initial test module, where users set up their initial hypothesis. After the initial hypothesis is set up, the relevant table(s) and test(s) are rendered and conducted. The following module is the Test diagnostics module, within which diagnostic tests such as the F-test and/or hypothesis analysis is done. Using Redhyte up till this point in the framework may already be sufficient for some users, as the hypothesis and test that they were interested in would be sufficiently addressed by the Initial test and the Test diagnostics module.

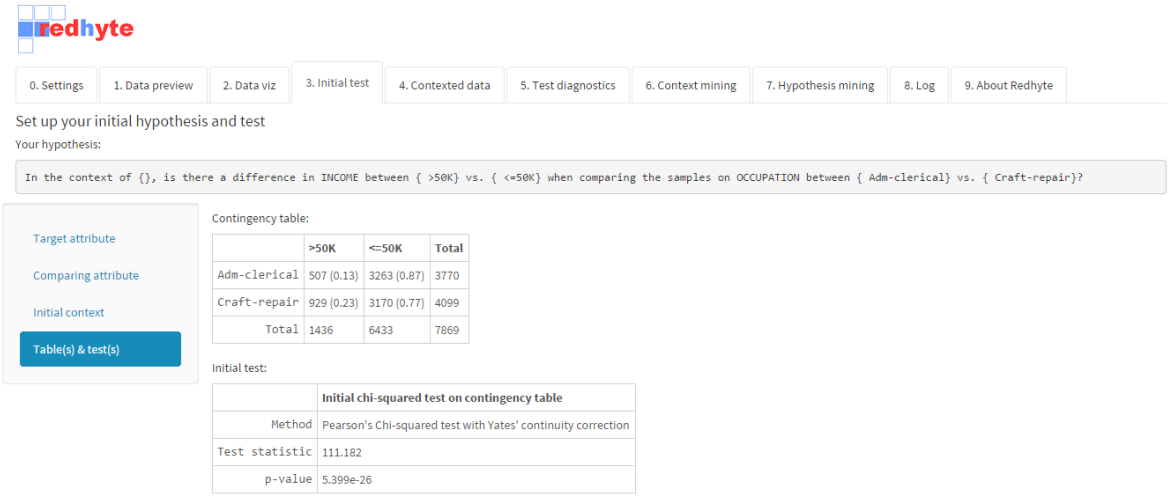
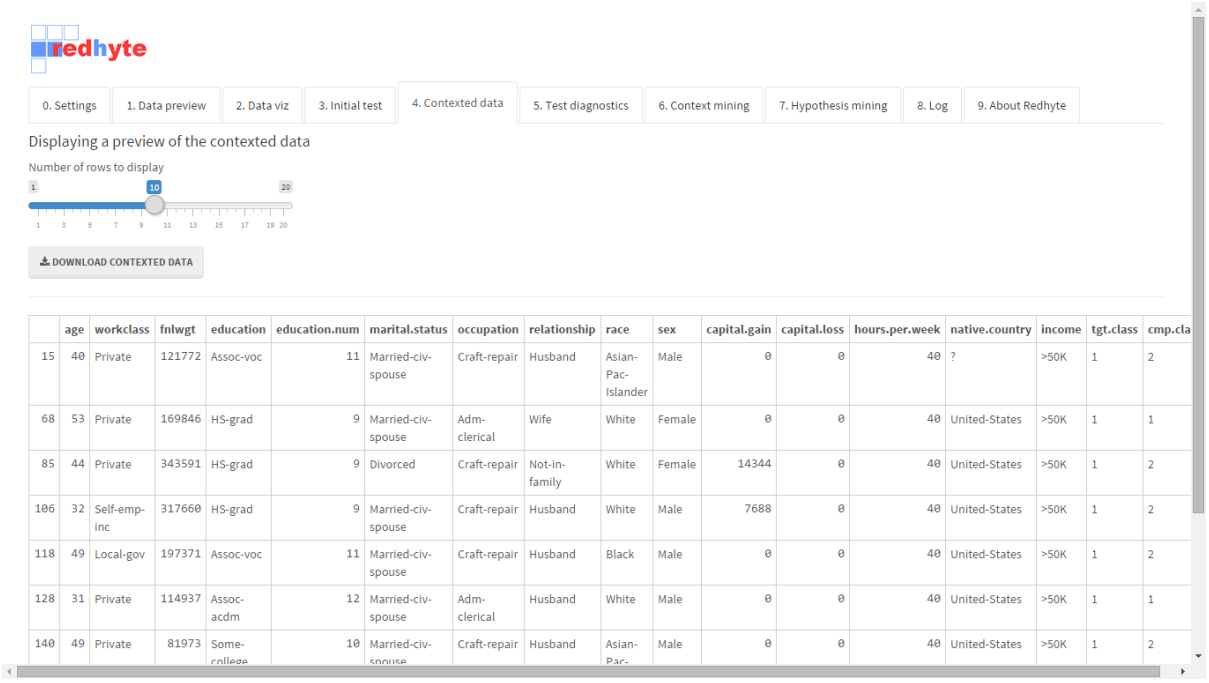



Figure 16: Screenshot of the Initial test module in Redhyte

The Contexted data tab allows users to have a quick look at the subset of the original input data that is relevant to the initial hypothesis ("contexted" simply means the addition of context items into a hypothesis, which makes the hypothesis more specific and the underlying dataset relevant to the hypothesis smaller). Furthermore, the less programming-savvy data analyst may make use of the Initial test module to do some simple subsetting of the original data, and download the data subset from the Contexted data tab for analysis in another platform or software.




| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country | income | tgt.class | cmp.cla |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 40 | Private | 121772 | Assoc-voc | 11 | Married-civ-spouse | Craft-repair | Husband | Asian-Pac-Islander | Male | 0 | 0 | 40 | ? | >50K | 1 | 2 |
| 68 | 53 | Private | 169846 | HS-grad | 9 | Married-civ-spouse | Adm-clerical | Wife | White | Female | 0 | 0 | 40 | United-States | >50K | 1 | 1 |
| 85 | 44 | Private | 343591 | HS-grad | 9 | Divorced | Craft-repair | Not-in-family | White | Female | 14344 | 0 | 40 | United-States | >50K | 1 | 2 |
| 106 | 32 | Self-emp-inc | 317660 | HS-grad | 9 | Married-civ-spouse | Craft-repair | Husband | White | Male | 7688 | 0 | 40 | United-States | >50K | 1 | 2 |
| 118 | 49 | Local-gov | 197371 | Assoc-voc | 11 | Married-civ-spouse | Craft-repair | Husband | Black | Male | 0 | 0 | 40 | United-States | >50K | 1 | 2 |
| 128 | 31 | Private | 114937 | Assoc-acdm | 12 | Married-civ-spouse | Adm-clerical | Husband | White | Male | 0 | 0 | 40 | United-States | >50K | 1 | 1 |
| 140 | 49 | Private | 81973 | Some-college | 10 | Married-civ-spouse | Craft-repair | Husband | Asian-Pac- | Male | 0 | 0 | 40 | United-States | >50K | 1 | 2 |

Figure 17: Screenshot of the Contexted data tab in Redhyte

### 3.2.4 Context mining

The Context mining module first allows users to remove attributes that should not be included in the context mining procedure, e.g. duplicated, redundant, or irrelevant attributes. After context mining is completed, the confusion matrices of the classification models, a list of mined context attributes, and variable importance plots (to be elaborated later) of the models are rendered. Redhyte also allows users to get a quick glance of the class distributions of the mined context attributes, with respect to the initial hypothesis.

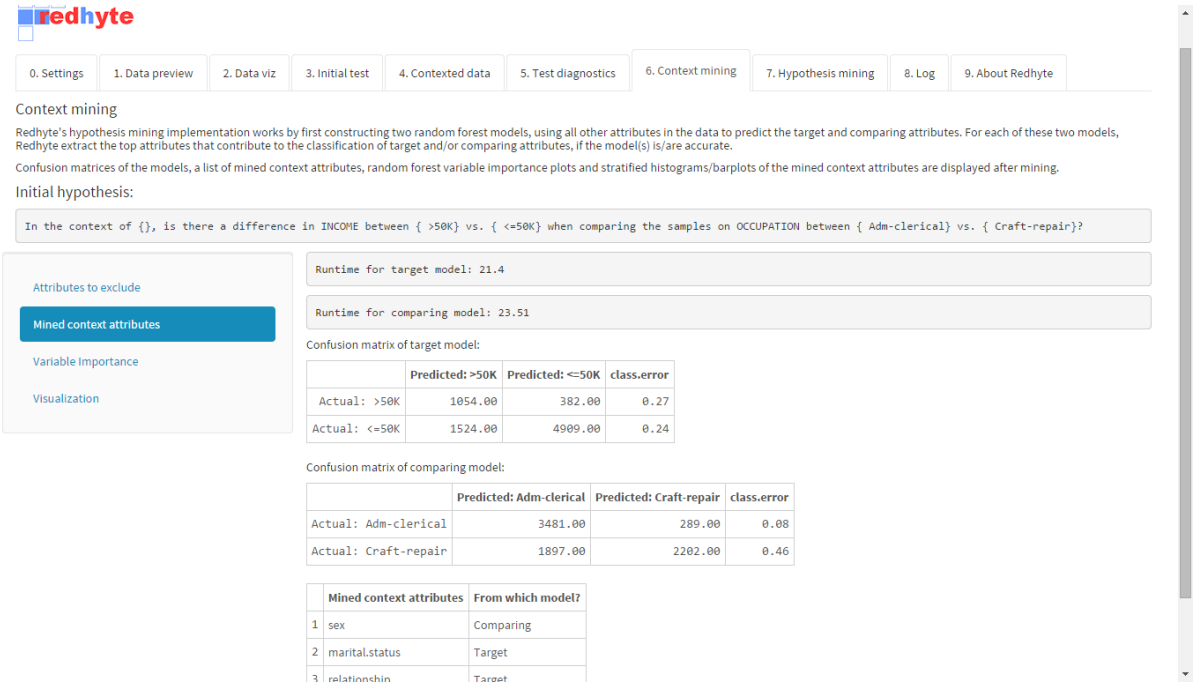
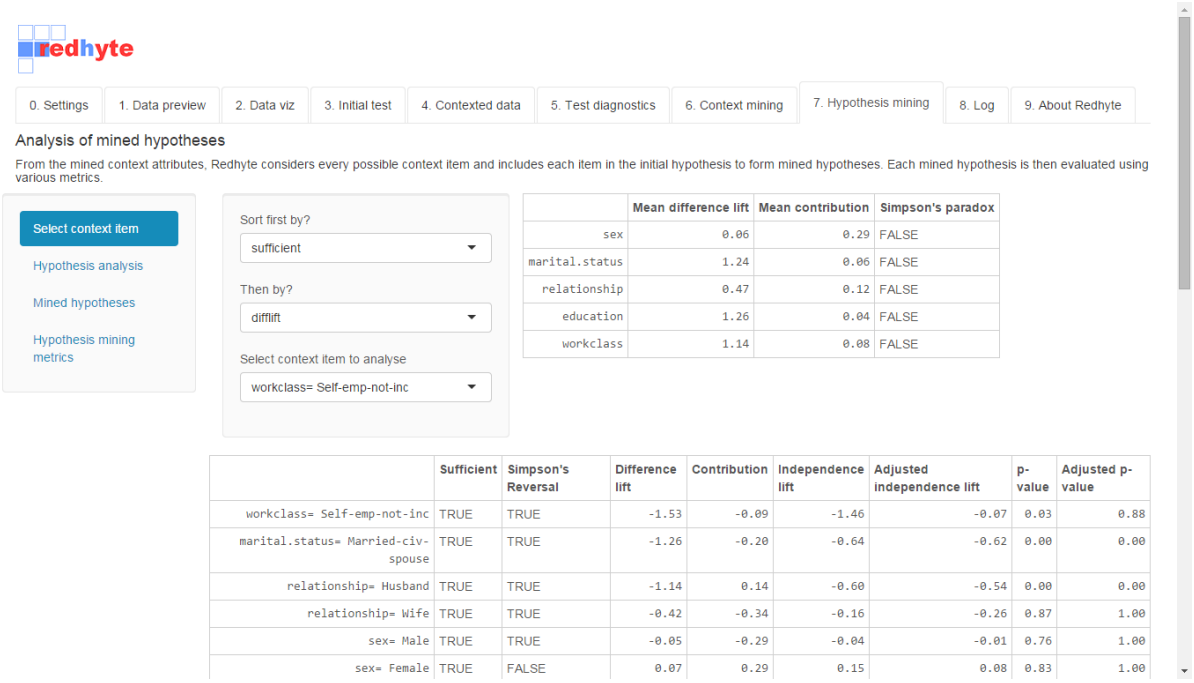



Figure 18: Screenshot of the Context mining module in Redhyte

### 3.2.5 Hypothesis mining

Using the mined context attributes, Redhyte generates a list of mined hypotheses, suitably scored by the hypothesis mining metrics. In the Hypothesis mining module, users can rank the mined hypotheses according to the hypothesis mining metrics, such as the difference lift and the independence lift. Users can also quickly identify the mined hypotheses in which Simpson's Reversals occurred; this cannot be easily done without the use of statistical programming in data analysis.




| | Mean difference lift | Mean contribution | Simpson's paradox |
|---|---|---|---|
| sex | 0.06 | 0.29 | FALSE |
| marital.status | 1.24 | 0.06 | FALSE |
| relationship | 0.47 | 0.12 | FALSE |
| education | 1.26 | 0.04 | FALSE |
| workclass | 1.14 | 0.08 | FALSE |

| | Sufficient | Simpson's Reversal | Difference lift | Contribution | Independence lift | Adjusted independence lift | p-value | Adjusted p-value |
|---|---|---|---|---|---|---|---|---|
| workclass= Self-emp-not-inc | TRUE | TRUE | -1.53 | -0.09 | -1.46 | -0.07 | 0.03 | 0.88 |
| marital.status= Married-civ-spouse | TRUE | TRUE | -1.26 | -0.20 | -0.64 | -0.62 | 0.00 | 0.00 |
| relationship= Husband | TRUE | TRUE | -1.14 | 0.14 | -0.60 | -0.54 | 0.00 | 0.00 |
| relationship= Wife | TRUE | TRUE | -0.42 | -0.34 | -0.16 | -0.26 | 0.87 | 1.00 |
| sex= Male | TRUE | TRUE | -0.05 | -0.29 | -0.04 | -0.01 | 0.76 | 1.00 |
| sex= Female | TRUE | FALSE | 0.07 | 0.29 | 0.15 | 0.08 | 0.83 | 1.00 |

Figure 19: Screenshot of the Hypothesis mining module in Redhyte

Based on the hypothesis mining metrics, users can select mined hypotheses that are deemed interesting for analysis. A comparison between the initial and the selected mined hypotheses (Figure 20) allows the user to quickly identify the rationale behind the (lack of) interestingness of the selected mined hypothesis, be it directed shrinkage (refer to section 4.5.1) or insufficient support. Finally, advanced users may wish to investigate the behaviour of the hypothesis mining metrics, using scatterplots of the various metrics.

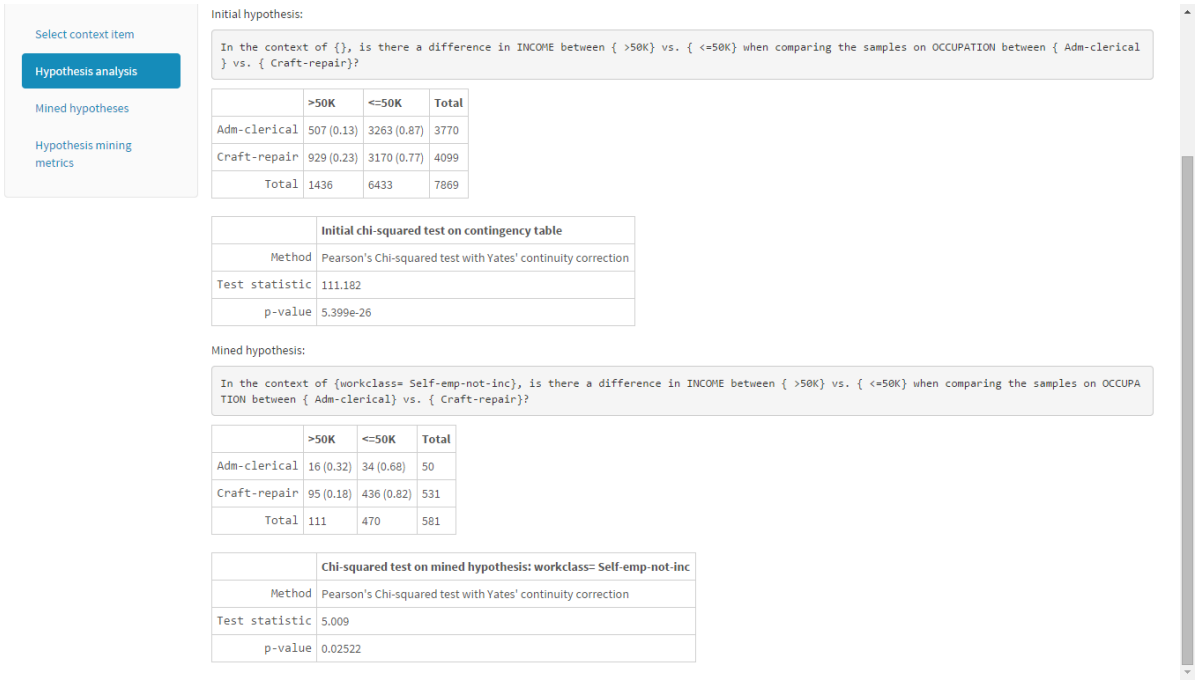



Figure 20: Screenshot of the Hypothesis analysis functionality in the Hypothesis mining module

### 3.2.6 Log

Finally, the Log documents all settings used in a particular analysis session, and allows users to quickly profile the analysis session. The log can be downloaded as a .csv file, and shared amongst collaborators for reproducibility of hypothesis mining and analysis results.

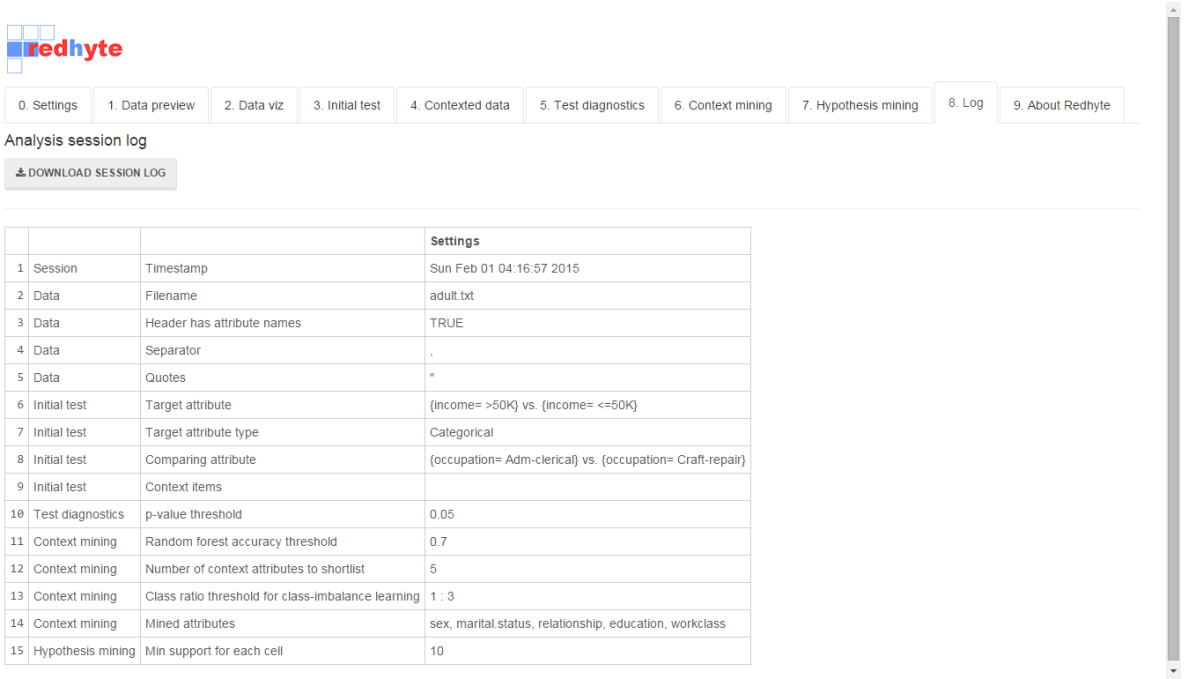



Figure 21: Screenshot of the Log tab in Redhyte

### 3.3 Interestingness of mined hypotheses

In this section, we give an illustration on how Redhyte can be used to lead the user to interesting mined hypotheses, using the UC Berkeley admission and the *adult* dataset. We use the following hypotheses as illustrations:

Table 22: Hypotheses from the UC Berkeley admission and the *adult* dataset

| | |
|---|---|
| Hypothesis A | In the context of {}, is there a difference in ADMIT between {Admitted} vs. {Rejected} when comparing the samples on GENDER between {Male} vs. {Female}? |
| Hypothesis B | In the context of {race= White}, is there a difference in INCOME between { >50K} vs. { <=50K} when comparing the samples on OCCUPATION between { Adm-clerical} vs. { Craft-repair}? |

*UC Berkeley admission* Based on hypothesis A, the initial test suggests the relationship between admission numbers and gender is significant ($p < 0.05$), with the males being more likely to be admitted into the university than females:

Table 23: Contingency table of Hypothesis I

| | **Admitted** | **Rejected** | **Total** |
|---|---|---|---|
| **Males** | 1198 (44.5%) | 1493 (55.5%) | 2691 |
| **Females** | 557 (30.4%) | 1278 (69.6%) | 1835 |
| **Total** | 1755 | 2771 | 4526 |

However, stratifying by various departments (departments A to F) gives different conclusions. In particular, inserting the context item {Dept = A} gives the following contingency table, with a p-value less than 0.05:

Table 24: Contingency table of mined hypothesis with {Dept = A}

| **{Dept = A}** | **Admitted** | **Rejected** | **Total** |
|---|---|---|---|
| **Males** | 512 (62.1%) | 313 (37.9%) | 825 |
| **Females** | 89 (82.4%) | 19 (17.6%) | 108 |
| **Total** | 601 | 332 | 933 |

Clearly, a Simpson's Reversal has taken place. When considering the admission rate of a particular department A, females are favoured for admission than males. This is contrary to the conclusion given by the initial test. Setting up the above hypothesis in Redhyte will allow the user to easily arrive at Table 24 in a matter of seconds.

*Adult* Based on hypothesis B in Table 22, the initial test suggests that the relationship between income and occupation is significant ($p < 0.05$), with white administrative clerks earning more than white craft repairers, as shown in Table 25.

Table 25: Contingency table of Hypothesis III

| | **Income > 50K** | **Income <= 50K** | **Total** |
|---|---|---|---|
| **Administrative clerks** | 439 (14.2%) | 2645 (85.8%) | 3084 |
| **Craft repairers** | 844 (22.8%) | 2850 (77.2%) | 3694 |
| **Total** | 1283 | 5495 | 6778 |

Using the default settings, Redhyte identifies five mined context attributes after context mining, namely sex, relationship, workclass, education, and education.num. In particular, considering the context items {Sex = Male}, {Sex = Female} and {Workclass = Self-emp-not-inc} leaves us with the following contingency tables:

Table 26: Contingency table of mined hypothesis with {Sex = Male}

| **{Sex = Male}** | **Income > 50K** | **Income <= 50K** | **Total** |
|---|---|---|---|
| **Administrative clerks** | 251 (24.2%) | 787 (75.8%) | 1038 |
| **Craft repairers** | 829 (23.5%) | 2695 (76.5%) | 3524 |
| **Total** | 1080 | 3482 | 4562 |

Table 27: Contingency table of mined hypothesis with {Sex = Female}

| **{Sex = Female}** | **Income > 50K** | **Income <= 50K** | **Total** |
|---|---|---|---|
| **Administrative clerks** | 188 (9.2%) | 1858 (90.8%) | 2046 |
| **Craft repairers** | 15 (8.8%) | 155 (91.2%) | 170 |
| **Total** | 203 | 2013 | 2216 |

The above illustrates an exact instance of a Simpson's Paradox, with both genders resulting in reversals of trends. This is also an example hypothesis mined by Liu *et al*'s hypothesis mining system, used as a case study in Liu *et al*.

Table 28: Contingency table of mined hypothesis with {Workclass = Self-emp-not-inc}

| **{Workclass = Self-emp-not-inc}** | **Income > 50K** | **Income <= 50K** | **Total** |
|---|---|---|---|
| **Administrative clerks** | 16 (34.8%) | 30 (65.2%) | 46 |
| **Craft repairers** | 90 (18%) | 409 (82%) | 499 |
| **Total** | 106 | 439 | 545 |

The hypothesis mining metrics (each hypothesis mining metric is designed to capture a specific aspect of hypothesis interestingness, to be discussed in section 4.5.1) evaluated on these three items are as follows:

Table 29: Hypothesis mining metrics evaluated for the selected context items

| Context items | Difference lift | Contribution | Independence lift | Adjusted independence lift | p-value |
|---|---|---|---|---|---|
| **{Sex = Male}** | -0.08 | -0.31 | -0.06 | -0.02 | 0.69 |
| **{Sex = Female}** | -0.04 | 0.31 | -0.09 | -0.05 | 0.98 |
| **{Workclass = Self-emp-not-inc}** | -1.94 | -0.11 | -1.89 | -0.05 | 0.01 |

Based on Hypothesis B, the default settings in Redhyte is used to illustrate the above, and to generate 27 other mined hypotheses, suitably scored and ranked using the hypothesis mining metrics, for the user to inspect. The user is also able to alter the settings in Redhyte – for instance, increasing the number of context attributes to mine for, to suit analysis purposes. All of these analysis results are entirely reproducible, with the help of the session log documentation.

### 3.4 Computational performance

There are two main ways Redhyte can be put to use: the user may choose to install R on a personal computer, and import Redhyte's GitHub repository to use Redhyte for analysis locally. Alternatively, Redhyte has been deployed as an R shiny application at shinyapps.io, for free, and can be accessed easily via the web browser. There is a limit to the upload size of the input data, which is at 20MB. Computations are significantly quicker on the server. We made comparisons on computation speeds on the server, versus a personal computer with AMD 2.2GHz Quad-Core processor with 64-bit operating system and 3.74GB usable RAM. The computational bottleneck of Redhyte comes from the construction of the context mining models (specifically, random forest models; refer to section 4.4.1) during context mining: as an example, for an initial hypothesis with about 7,800 samples, context mining takes approximately 45 seconds on the said personal computer, and 16 seconds on the server. Other than context mining, the reminder of the background computations and data manipulations in Redhyte is relatively lightweight and do not pose any problems, as long as they do not require excessive amounts of computational resources that the backend server at shinyapps.io has allocated to Redhyte for free. The following tabulates the performance comparison on various datasets, hypotheses, and hardware.

Table 30: Initial hypotheses used to evaluate Redhyte's computational performance

| | Dataset | Initial hypothesis |
|---|---|---|
| **I** | *UC Berkeley* | In the context of {}, is there a difference in ADMIT between {Admitted} vs. {Rejected} when comparing the samples on GENDER between {Male} vs. {Female}? |
| **II** | *Adult* | In the context of {}, is there a difference in INCOME between { >50K} vs. { <=50K} when comparing the samples on OCCUPATION between { Adm-clerical} vs. { Craft-repair}? |
| **III** | *Adult* | In the context of {race= White}, is there a difference in INCOME between { >50K} vs. { <=50K} when comparing the samples on OCCUPATION between { Adm-clerical} vs. { Craft-repair}? |
| **IV** | *Adult* | In the context of {}, is there a difference in AGE when comparing the samples on WORKCLASS between { Federal-gov & State-gov} vs. { Private}? |
| **V** | *Adult* | In the context of {sex= Male}, is there a difference in AGE when comparing the samples on WORKCLASS between { Federal-gov & State-gov} vs. { Private}? |
| **VI** | *Mushroom* | In the context of {}, is there a difference in CLASS between {POISONOUS} vs. {EDIBLE} when comparing the samples on BRUISES between {BRUISES} vs. {NO}? |
| **VII** | *Arrhythmia* | In the context of {}, is there is a difference in CLASS between {NORMAL} VS. {DISEASED} when comparing the samples on GENDER between {MALE} vs. {FEMALE}? |

Table 31: Comparison of Redhyte's computational performance across datasets, initial hypotheses, and hardware

| Initial Hypothesis | Dataset | Number of samples in initial hypothesis | Number of attributes | Context mining runtime on shinyapps.io server | Context mining runtime on personal computer |
|---|---|---|---|---|---|
| **I** | *UC Berkeley* | 4526 | 3 | 2.364secs | 4.48secs |
| **II** | *Adult* | 7,869 | 15 | 16.442secs | 44.91secs |
| **III** | *Adult* | 6,778 | 15 | 13.078secs | 34.29secs |
| **IV** | *Adult* | 24,954 | 15 | *Exceeded server limits* | 199.89secs |
| **V** | *Adult* | 16,398 | 15 | *Exceeded server limits* | 110.64secs |
| **VI** | *Mushroom* | 8,416 | 23 | 9.578secs | 32.67secs |
| **VII** | *Arrhythmia* | 452 | 280 | 4.193secs | 10.72secs |

# Discussion

## 4.1 Hypothesis mining

Redhyte is a hypothesis mining system that utilises the user's initial, domain knowledge-driven hypothesis to search for relevant and interesting hypothesis. Redhyte was motivated by the lacklustre practicality of hypothesis testing in the current Big Data context. In circumstances where the number of attributes in the dataset is too large for the data analyst to purposefully examine and scrutinize, hypothesis testing is often entirely replaced by more wholesome approaches, such as data mining. This is for good measure, as not only is formulating and testing a small hypothesis in a large dataset wasteful, it is also flawed. Data mining techniques are able to accommodate large number of attributes and use them towards some meaningful statistical task or objective, such as classifying or clustering. This use of data in aggregation, while wholesome, is not nearly as intuitive as making direct comparisons between subpopulations of interest, as in hypothesis testing. Motivated as such, Redhyte was designed to be able make full use of entire datasets while remaining within the hypothesis testing framework. Based on the user's initial hypothesis of interest, Redhyte searches for interesting hypotheses for the user to consider and inspect. Redhyte also conducts diagnostics and analyses to sufficiently address the initial hypothesis.

The actions required from the user are to upload the dataset and set up the initial hypothesis. This is in contrast to Liu *et al*'s hypothesis mining system, whereby the mined hypotheses are not driven by initial domain knowledge input. In other words, the mined hypotheses generated by Liu *et al*'s system are definitive and entirely data-driven – for a given dataset and a set of mining parameters, the system generates a specific list of mined hypotheses. In contrast, Redhyte is both domain knowledge- and data-driven – Redhyte first takes into consideration the user's initial hypothesis, and mines for hypotheses that are relevant to it.

There are several distinctions between both approaches. For example, Redhyte can never mine for a hypothesis that is not related to the initial hypothesis. If the user's initial hypothesis is one that is well-formulated, then Redhyte's mined hypotheses would be of value to the user. On the contrary, Liu *et al*'s system does not allow the user to stipulate any initial hypothesis. Hence, the hypotheses mined by Liu *et al*'s system could possibly contain spurious ones – one example of such a hypothesis could be the comparison of the number of males and females who are husbands or wives. Without any domain knowledge input from the user, the system is unable to discard such hypotheses.

Inevitably, Redhyte presumes that the user is sufficiently knowledgeable to formulate an initial hypothesis that is meaningful and relevant in subject matter in the first place. Without an initial hypothesis that is relevant in subject matter, Redhyte would be unable to mine for any interesting hypothesis. On the other hand, even without a knowledgeable user, Liu *et al*'s hypothesis mining system would function just as well, with the mined hypotheses allowing the user to get a good understanding of the input dataset.

### 4.2 Initial test

Redhyte is divided into mainly four modules, namely the initial test, test diagnostics, context mining, and the mined hypothesis formulation and scoring modules. The initial hypothesis is assessed by either the t-test or the $\chi 2$ test in Redhyte, depending on whether the target attribute is numerical or categorical. The t-test is used to make comparison between two subpopulations. Amongst the various variants and uses of the t-test, the most typical and familiar use of the t-test is to compare the means of the target attribute across two subpopulations, and that is how Redhyte uses the t-test in the initial test set-up. The t-test asserts in its null hypothesis that the means of the target attribute in the subpopulations are

not different from each other. As a parametric test, the t-test requires certain assumptions to hold true. These assumptions include i) the distributions of the numerical target attribute in the subpopulations under comparison are approximately normal, ii) the target attribute has roughly equal variances in the subpopulations, and iii) data of target attribute is independently and identically sampled in the subpopulations (i.i.d.). Typically, of these three assumptions, only the first two can be assessed rigorously using other statistical tests, while a conclusive statistical test is not available to assess the third. These assumptions are assessed in the test diagnostics module in Redhyte.

For categorical target attributes, the initial hypothesis is assessed by the $\chi 2$ test. The $\chi 2$ test can be considered to be a non-parametric test, and asserts that the target and comparing attributes are not associated to each other in its null hypothesis. The $\chi 2$ test is often accompanied by contingency tables, such as the one shown in Table 5. Redhyte also uses the $\chi 2$ test to assess relationship between a numerical target attribute and the comparing attribute, with a binary attribute corresponding to a mean discretization on the target attribute.

### 4.3 Test diagnostics

Subsequently, Redhyte proceeds to the test diagnostic module, whereby a check of test assumptions and/or hypothesis analysis is done. For the t-test, Redhyte uses the Shapiro-Wilk test and the F-test to test for normality and homoscedasticity. If either of these tests fails, Redhyte then uses the Mann-Whitney test to assess the hypothesis. These tests are implemented in the system backend as R functions, and for computational reasons, the Shapiro-Wilk test implementation in R is unable to accommodate subpopulation sizes of more than 5,000 samples. If need be, the Shapiro-Wilk test in Redhyte could be re-implemented. For the collapsed $\chi 2$ test, the initial hypothesis is analysed, to identify the $\chi 2$

top contributor. We refer the reader to section 2.1.4 to understand the utility of identifying the $\chi 2$ top contributor.

The final part of the test diagnostics module is the CMH test, or the Cochran–Mantel–Haenszel test. For a given categorical attribute that is not the target or the comparing attribute, the test considers all of its classes and stratifies the contexted data accordingly, to form *k* 2x2 tables. With a null hypothesis that the target and comparing attributes are conditionally independent across the *k* tables, rejection of the null hypothesis implies that the stratifying categorical attribute could potentially confound the initial hypothesis. In Redhyte, if any of the *k* tables have any cell support of zero after stratification, the attribute is flagged to be inadequate for the CMH test to be conducted. The goal of this test is to search for potential confounders to the initial hypothesis, and in some sense, is aligned with the goal of context mining. We make a comparison of the CMH test with context mining later in the section.

### 4.4 Context mining

After the initial test and test diagnostics, Redhyte proceeds to mine for potentially relevant context attributes. This is done using classification techniques from data mining. Given a categorical response attribute to predict, classification models make use of other attributes in the dataset to sieve out some form of empirical structure to facilitate prediction and classification. For example, in a decision tree classifier predicting resting heart rate, the classifier may learn empirically from the data that smokers generally have a resting heart rate being above the mean of the dataset, and use that rule as a basis of classification.

In Redhyte, two separate classification models are used to predict, using all other attributes, the target and the comparing attribute. We call these models, for simplicity, the target and comparing model respectively. The rationale stems from the fact that if some attribute *A* in the dataset is related with the target or the comparing attribute, then the mined hypothesis formed by using *A* should be interesting to consider. One way to identify such relationships in the data, without using an iterative approach, is to use a classification model. The classification model used in Redhyte is the random forest model (Breiman, 1996a, b, 2001). The selection of a model suitable for context mining is largely guided by two requirements: attribute selection and accuracy, though there are other points of consideration. We return to these points later.

Context attributes are mined in Redhyte in the following manner: the classification accuracies of the target and the comparing model are computed. If neither model has accuracy above the user-defined classification accuracy threshold, then context mining stops – we say that there are no context attributes to be mined. If either model has an accuracy above the threshold, then Redhyte takes the top *k* attributes from that model (*k* is defined by the user; ranking of the attributes is done within the model, to be elaborated further), and call them the mined context attributes. If both models are accurate, Redhyte considers the top *k* attributes from the intersection set of the top attributes from both models as the mined context attributes. Therefore, a mined context attribute in Redhyte, by definition, contributes significantly to the classification of either the target or comparing attribute, or both.

### 4.4.1 Random forest

As the name suggests, random forest models basically contain an ensemble of decision tree models, with each decision tree predicting the same categorical response attribute (Breiman,

1996a, b, 2001). A simple ramification of such an ensemble learning technique is that it is possible to get a majority vote on the predicted classification over the numerous decision trees, thereby increasing the classification accuracy of the "forest" as a whole. To construct individual decision trees in the ensemble, random forest models use a sampling procedure known as bootstrap aggregating, otherwise known as bagging. Each decision tree is constructed using a "bag" of samples randomly drawn from the dataset with replacement. Such a sampling procedure allows random forest models to be robust to noise: if the dataset contains some noise or mistakes, especially in the response attribute, it can be shown that bagging ensures that the number of bags with more noise will be less than that of bags with less noise (Breiman, 1996a, b). In this way, random forest models attain a certain degree of robustness to noise. In addition, to de-correlate the individual trees in the random forest model, the model uses a technique known as random subspace sampling (Bryll, 2003) – each split on each tree is only decided based on a subset of, typically, $\sqrt{p}$ attributes, where $p$ is the total number of attributes in the dataset. This is to ensure that the individual trees are less similar to each other, since they are splitting at each node using possibly different attributes. Having less similar and more independent trees in the forest increases classification performance, due to a phenomenon known as the wisdom of crowds (Rokach, 2010). For more technical details on random forest models, refer to Breiman (1996a, b, 2001).

### 4.4.2 Attribute selection

Redhyte's use of classification models demands some form of scoring of attributes within the model, in order to rank the attributes based on how well an attribute contributes to the classification. The better an attribute can classify the target or the comparing attribute, the more interesting it might be to be considered as a context attribute. This is essentially equivalent to a means for attribute selection in general data mining tasks (Saeys *et al*, 2007),

and random forest models are able to do exactly that. Inherent in random forest models are two ways using which attributes can be ranked: the mean decrease in Gini (MDG) impurity and the mean decrease in accuracy (MDA). In Redhyte, the MDA is used as the criterion, or variable importance measure, for attribute selection. MDA works in the following way: to assess the importance of a given attribute *A*, the values of *A* are first randomly permutated in the dataset. If a model constructed on the original dataset has roughly the same accuracy as one constructed on the dataset with *A* randomly permutated, then we say that *A* is not an important attribute. This is likewise for the converse. Random forest models compute MDA measures for all attributes in the dataset. Both MDA and MDG, while numerically different, are known to produce variable importance in which the ranking of the attributes are very similar (Hastie *et al*, 2009; Kawakubo and Yoshida, 2012). Classification models such as the naïve Bayesian classifiers (Russell and Norvig, 1995) and artificial neural networks (Wang, 2003), either do not generate such scoring of attributes, or generate scoring based on statistically transformed attributes (e.g. linear combinations of attributes in principal components analysis), and therefore are not suited to be context mining models in Redhyte. Redhyte renders variable importance dotplots for both measures in the context mining module.

### 4.4.3 Classification accuracy and class-imbalance learning

To evaluate and visualise the accuracy of a classification model, one may use a confusion matrix, as such:

Table 32: An example confusion matrix

| | **Predicted as positive** | **Predicted as negative** |
|---|---|---|
| **Actual positive** | True positives, TP | False negatives, FN |
| **Actual negative** | False positives, FP | True negatives, TN |

In the confusion matrix above, the classifier is trying to predict a binary response attribute that can either be positive or negative. The accuracy of this classifier model is evaluated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Numerous empirical investigations conducted across various domains suggest that the random forest is amongst the top classification models (Svetnik *et al*, 2003; Caruana and Niculescu-Mizil, 2006; Caruana *et al*, 2008; Brown and Mues, 2012; Gupta *et al*, 2012). In particular, Fernandez-Delgado *et al* (2014) showed that random forest models outperform many different classification models such as linear discriminant analysis, naïve Bayesian classifiers, and decision trees in terms of classification accuracy.

Mining for interesting context attributes requires high classification accuracy. A well-known issue in classification problems is learning with empirical class-imbalance in the data (He and Garcia, 2009; Kotsiantis *et al*, 2006). Consider the following classifier performances:

Table 33: Example confusion matrices and their associated accuracies

| Classifier | TP | TN | FP | FN | Accuracy |
|---|---|---|---|---|---|
| **A** | 30 | 100 | 50 | 20 | 65% |
| **B** | 0 | 150 | 0 | 50 | 75% |

The data used in classifier B is clearly unbalanced – there are 3 times as many actual negatives than there are positives in B. Yet, by using a naïve prediction rule of always predicting negatives, the accuracy of B triumphs that of A, simply because of class-imbalance. This simple example shows that accuracy, as a performance measure for classification models, is unreliable when the data is imbalanced. There have been many proposals that serve to address class-imbalance, such as the geometric mean (He and Garcia, 2009; Weiss, 2004) and the F-measure (Kubat and Matwin, 1997; Akbani *et al*, 2004) as alternative performance metrics for classification models. Chawla *et al* suggested the synthetic minority over-sampling technique (SMOTE) to synthetically generate samples of

the minority class to correct for class-imbalance. In Redhyte, correction for class-imbalance learning is done by using the adjusted geometric mean ($AG_m$) (Batuwita and Palade, 2011):

$$AG_m = \frac{G_m + SP \cdot N_n}{1 + N_n},$$

where $G_m$ refers to the geometric mean accuracy of the model, $SP$ the specificity of the model, and $N_n$ the proportion of samples that belong to the majority class.

One reason why class-imbalance could exist, other than the data being inherently imbalanced, is the use of mean discretization in Redhyte to discretize the target attribute. Using the mean as a discretization threshold often results in class-imbalance whenever the data has a skewed distribution that deviates far from the Gaussian Normal. This is because the mean, as a measure of central tendency, is not as robust against outliers, as compared to other measures of central tendency, such as the median. While we could have implemented other measures of central tendency or a data-driven discretization algorithm (e.g. entropy-based discretization algorithms (Fayyad and Irani, 1993)), Redhyte uses mean discretization mainly because i) Redhyte uses the t-test on means as one of its two initial tests, and ii) mean discretization is intuitive and easily understood. Also, we note that random forest models are known to tolerate levels of class-imbalance better than most classification models (Brown and Mues, 2012; Chen *et al*, 2004).

### 4.4.4 Context mining versus CMH test

Context mining seeks out attributes that could potentially confound the initial hypothesis. At first glance, it may seem that the CMH test is able to serve as a model from which context attributes can be mined. However, there are several limitations that are inherent to the CMH test. Firstly, if the trends observed are in opposite directions across some of the *k* tables, then the CMH test is not an appropriate test – the test works well only if the trends across all of the

tables are in the same direction, and are comparable in size (Agresti, 1996). Furthermore, the null hypothesis of the CMH test is that of conditional independence, i.e. all conditional odds ratio in the *k* tables are equal to 1. Therefore, for an initial hypothesis that is not marginally independent; that is, ignoring any third attribute stratification, the trend observed between the target and the comparing attributes is significant, the CMH test could be significant for attributes that do not cofound the initial hypothesis.

Moreover, the CMH test requires the use of *k* contingency tables. Loosely speaking, as *k* increases, the support for each table decreases, hence limiting the statistical significance of the test. In addition, with large number of attributes in the dataset, using the CMH test for each and every attribute is subjected to multiple testing problems. These problems are also applicable to other statistical tests similar to the CMH test, such as the Breslow-Day test (Breslow and Day, 1980) or the DerSimonian-Laird test (DerSimonian and Laird, 1986), but are not relevant for classification models.

### 4.4.5 Equivalent models/methods in context mining

In principle, context mining can be done using any form of classification model, such as the regression-based logistic regression, or even using correlation measures – if an attribute is correlated with the target or the comparing attribute, then it might be worthwhile to consider the attribute as a mined context attribute. To rationalize the use of classification models in context mining, we assume that the input dataset given to Redhyte is entirely generic and arbitrary; that is, we may have various issues such as i) multicollinearity, or correlation between attributes, ii) nonlinearity between the target and comparing attribute, and the other attributes in the dataset, and iii) class-imbalance. Given these unforeseen but plausible characteristics of the input dataset, generalized linear models and correlation measures may

not be the best option. Furthermore, correlation measures consider each attribute separately and are hardly feasible when the number of attributes is too large. That leaves us with the use of classification models for context mining. We chose to implement the random forest model as our context mining model of choice in Redhyte, as its empirical performance has been documented and well-received by many (Svetnik *et al*, 2003; Caruana and Niculescu-Mizil, 2006; Caruana *et al*, 2008; Brown and Mues, 2012; Gupta *et al*, 2012; Fernandez-Delgado *et al*, 2014). In addition, the random forest model allows for easy attribute selection due to the nature of the model (Svetnik *et al*, 2003; Genuer *et al*, 2010). In particular, using the permutation-based MDA as a variable importance measure exactly accords to the intuition of context mining: recall that to compute the MDA of an attribute *A*, the values of *A* is randomly permutated in the dataset. If such a permutation affects the accuracy of the model, then *A* is an important attribute and is related to either the target and/or comparing attribute. Finally, random forest models do not require any prior data transformation or normalization, nor do they require disparate training and testing datasets for cross-validation – classification accuracies can be computed using "out-of-bag" estimates (Breiman, 1996a, b, 2001). These properties allow random forest models to be ideal models for context mining in Redhyte.

### 4.5 Mined hypothesis formulation, scoring and ranking

After context mining, Redhyte uses the list of mined context attributes and construct mined hypotheses, by considering every possible context item in the mined context attributes: for example, if gender is a mined context attribute, then the mined context items that Redhyte considers are {Gender = M} and {Gender = F}. Each of these context items are then added into the initial hypothesis, which shrinks the subpopulations of the hypothesis. We call the resultant hypotheses mined hypotheses. Only mined hypotheses with cell supports exceeding

the minimum cell support stipulated in the settings (default at 10) are considered to be worthy for inspection and analysis.

### 4.5.1 Hypothesis mining metrics

To evaluate the mined hypotheses and divert the user's attention to the most interesting collection of mined hypotheses, Redhyte uses four different hypothesis mining metrics to rank the mined hypotheses. Two of these metrics, the difference lift and the contribution, have been previously published in Liu *et al*. We developed and implemented two additional metrics in Redhyte, namely the independence lift and the adjusted independence lift, to improve the scoring and ranking of the mined hypotheses. The Appendix outlines the intuition, basis, and derivation of the various metrics. Each metric was designed to capture specific aspects of interestingness of the mined hypotheses: trend changes, relative support of mined hypotheses, and manner of shrinkage. The following summarizes the metrics and the aspects of interestingness that each metric captures.

Table 34: Summary of the various hypothesis mining metrics used in Redhyte

| Metrics | Trend changes | Relative support | Shrinkage manner | Remarks |
|---|---|---|---|---|
| **Difference lift** | ✓ | | | Captures changes in trends and proportions after addition of mined context item |
| **Contribution** | | ✓ | | Allow mined hypotheses to be scored according to the relative support of the mined hypotheses |
| **Independence lift** | ✓ | ✓ | | Evaluate mined hypotheses based on the above two aspects of interestingness simultaneously |
| **Adjusted independence lift** | ✓ | | ✓ | To capture changes in trends and proportions, and the manner of shrinkage (directed or undirected) |

*Trend changes* Trend changes refer to changes in proportions or trends, be it amplifications, weakening, or reversals, when a mined context item is added into the initial hypothesis.

Trend changes are captured by the difference lift, the independence lift, and the adjusted independence lift. Specifically, each of these metrics is proportional to change in trend, and has a property that if a Simpson's Reversal occurs, these metrics are numerically negative. This property gives us a simple and quick way to detect Simpson's Reversals.

*Relative support* If the mined hypothesis, after the addition of a mined context item, still retains a large support relative to that of the initial hypothesis, then intuitively the mined hypothesis could be more interesting to consider, as it still retains some form of generality. In contrast, if the subpopulations of the mined hypothesis shrink to a very small number of samples, then this mined hypothesis could be too specific, less useful, and hence less interesting. The contribution and independence lift of a context item is proportional to the relative support of the mined hypothesis formed by it.

*Shrinkage manner* When a mined context item is added into the initial hypothesis, the subpopulations of the resultant hypothesis shrink. Consider the following: the context item could, for example, shrink each cell count of the contingency table of the initial hypothesis, in a more or less uniform manner; perhaps subtracting very similar numbers of samples from each cell count. We call this *undirected shrinkage*. On the other hand, the context item may also shrink each count cell of the initial hypothesis in a more "directed" manner – if we consider the following contingency table, one example of *directed shrinkage* could be having $c_{11}$ reduced by a much larger extent, say halved, than $c_{12}$, $c_{21}$, and $c_{22}$, when a context item is added.

Table 35: Example contingency table

| | **High resting heart rate** | **Low resting heart rate** | **Total** |
|---|---|---|---|
| **Smokers** | $c_{11}$ = 43 (61.4%) | $c_{12}$ = 27 (38.6%) | 70 |
| **Never-smokers and quitters** | $c_{21}$ = 12 (21.4%) | $c_{22}$ = 44 (78.6%) | 56 |
| **Total** | 55 | 71 | 126 |

Intuitively, it is arguable that directed shrinkage may be more interesting, as directed shrinkage suggests an association of the context item with the initial hypothesis (we make a case for both directed and undirected shrinkage later in the section). The adjusted independence lift was designed to capture directed shrinkage – specifically, the adjusted independence lift of a context item is proportional to the extent of directed shrinkage that it induces. In the following section, we describe a scenario whereby directed shrinkage can be observed in reality.

### 4.6 What makes a hypothesis interesting?

While Redhyte aims to search for statistically significant and practically interesting hypotheses, throughout the development of Redhyte we note that it is not possible to objectively quantify the interestingness of a mined hypothesis. In addition to the trivial fact that interestingness itself is a subjective measure, in the following sections we give formal explanations as to why it is not possible to do so. First, we describe a concrete scenario that depicts directed shrinkage, and use it as a quick example in the following sections.

### 4.6.1 The Rhesus gene

Consider the following hypotheses on infant deaths and birth order:

Table 36: Contingency table of initial hypothesis on infant deaths

| $H_{initial}$ | **Infant died within 6 months of birth, $T_1$** | **Lived more than 6 months, $T_2$** |
|---|---|---|
| **First child, $C_1$** | $c_{11}$ $(p_1)$ | $c_{12}$ |
| **Not first child, $C_2$** | $c_{21}$ $(p_2)$ | $c_{22}$ |

Table 37: Contingency table of mined hypothesis with : $\{Rh^+ = True\}$

| $H_1{}^*$: $\{Rh^+ = True\}$ | **Infant died within 6 months of birth, $T_1$** | **Lived more than 6 months, $T_2$** |
|---|---|---|
| **First child, $C_1$** | $c'_{21}$ $(p'_2)$ | $c'_{22}$ |
| **Not first child, $C_2$** | $c'_{21}$ $(p'_2)$ | $c'_{22}$ |

Table 38: Contingency table of mined hypothesis with : $\{Rh^+ = False\}$

| $H_2{}^*$: $\{Rh^+ = False\}$ | **Infant died within 6 months of birth, $T_1$** | **Lived more than 6 months, $T_2$** |
|---|---|---|
| **First child, $C_1$** | $c''_{11}$ $(p''_1)$ | $c''_{12}$ |
| **Not first child, $C_2$** | $c''_{21}$ $(p''_2)$ | $c''_{22}$ |

The initial hypothesis concerns the association between infant deaths and birth order, while the mined context attribute is the presence of the Rhesus gene allele, $Rh^+$, in the biological parents. The Rhesus gene comes in two alleles, $Rh^+$, meaning the presence of the Rh antigen and $Rh^-$, the absence thereof. When an $Rh^-$ mother is impregnated by an $Rh^+$ father, proteins from the potentially $Rh^+$ fetus may enter the mother's bloodstream. This results in the sensitization of the mother's immune system to produce antibodies to attack $Rh^+$ blood cells. Therefore, in subsequent pregnancies these antibodies attack the fetal blood cells, resulting in hemolytic anemia of the fetus and, if no miscarriage occurs, high chance of death for the infant.

Assume in this hypothetical example that, given the rarity of the $Rh^+$ allele (Flegel, 2007), $H_{initial}$ displays no association between infant death and birth order. Understanding the mechanism of the Rh gene allows us reasonably state the following:

1. Adding $I = \{Rh^+ = True\}$ into $H_{initial}$ results in directed shrinkage of the subpopulations – specifically, the difference between $c_{22}$ and $c'_{22}$ should be disproportionately large.
2. $c'_{21}$ should be disproportionately larger than the other cell counts in ${H_1}^*$.
3. Given the rarity of the $Rh^+$ allele, the conclusion drawn from ${H_2}^*$ will be very similar to that of $H_{initial}$.

Using these intuitions, we can describe qualitatively the behavior of the difference lift and adjusted independence lift, with respect to the context item $I = \{Rh^+ = True\}$:

Table 39: Qualitative description of the difference lift and adjusted independence lift

| | |
|---|---|
| $DiffLift(I = \{Rh^+ = True\} \mid H_{initial})$ | Large (trend amplification) |
| $AdjustedIndpLift(I = \{Rh^+ = True\} \mid H_{initial})$ | Large (trend amplification and directed shrinkage) |

*Extent of domain knowledge* The interestingness of a hypothesis can never be objectively quantified as it is directly related to the current extents of domain knowledge. For instance, while both the difference lift and adjusted independence lift flag ${H_1}^*$ as an interesting hypothesis (Table 39), whether the above set of hypotheses is interesting solely depends on domain knowledge – if knowledge of the Rh gene is yet to be discovered, then the shrinkage of subpopulations directed by $I = \{Rh^+ = True\}$ will be a very groundbreaking finding. On the other hand, if the mechanism of Rh gene is well-understood, then the above set of hypotheses is essentially trivial.

*Undirected shrinkage versus directed shrinkage* When a context item $I$ is added into $H_{initial}$ to form $H^*$, the subpopulations of the hypothesis shrinks, as the resultant hypothesis

becomes more specific. In derivation of the adjusted independence lift (refer to Appendix C), we define the independence factor of a context item $i_I$, where $i_I$ measures the extent of association/independence $I$ has on the target attribute value $T_1$. Using the intuition that undirected shrinkage of subpopulations is less interesting than a directed shrinkage, the adjusted independence lift is corrected by $\left|1 - \frac{1}{i_I}\right|$ instead of $\frac{1}{i_I}$, as in independence lift. With this correction, the adjusted independence lift is able to capture directed shrinkages, such as the one illustrated by the $Rh^+$ allele.

Suppose $H_{initial}$ and $H^*$ are as follows:

Table 40: Example initial hypothesis

| $H_{initial}$ | Target attribute class $T_1$ | Target attribute class $T_2$ |
|---|---|---|
| **Comparing attribute class $C_1$** | 50 (0.71) | 20 |
| **Comparing attribute class $C_2$** | 60 (0.67) | 30 |

Table 41: Mined hypothesis with undirected shrinkage

| $I = \{A_{ctx} = v_{ctx}\}$ | Target attribute class $T'_1$ | Target attribute class $T'_2$ |
|---|---|---|
| **Comparing attribute class $C'_1$** | 45 (0.75) | 15 |
| **Comparing attribute class $C'_2$** | 55 (0.69) | 25 |

In the example above, the context item $I = \{A_{ctx} = v_{ctx}\}$ shrinks each cell of $H_{initial}$ by 5 samples. Given such a hypothesis, the adjusted independence lift interprets that $I$ and $T_1$ are independent, and hence be weighted down in interestingness, with $AdjustedIndpLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = 0.05$. On the other hand, $p'_1 > p_1, p'_2 > p_2$, and $DiffLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = 1.31$, i.e. the trend observed has been amplified by $I = \{A_{ctx} = v_{ctx}\}$. In this case the difference lift and the adjusted independence lift give contradictory conclusions, and without any domain knowledge input, it is not possible to ascertain the interestingness of the mined hypothesis.

Nonetheless, it is easy to give justification for both metrics, such as the following: if we were to make a discovery of a gene, say the Rh gene in the previous example, directed shrinkage would provide a clear indication on the relationship between $I$ and $T_1$. In this case, the adjusted independence lift would be facilitating the discovery of the gene. On the other hand, a marketing department's decision on whether to place an advertisement in the morning papers versus the evening papers (the context attribute) can solely be based on the difference lift on some target attribute associated with response or outreach, even with undirected shrinkage.

*Relative support of mined hypotheses* Both the contribution and the independence lift were designed to take into consideration the relative support of $H_{initial}$ and $H^*$. In particular, contribution and independence lift favours $H^*$ with larger relative support, using, for instance, the $\frac{n'}{n}$ coefficient in independence lift (refer to Appendix B), where $n'$is the support for the mined hypothesis while $n$ is that of the initial. However, the relationship between relative support and interestingness of mined hypotheses is hardly that straightforward. For instance, if $\frac{n'}{n}$ is close to 1, then adding the context item does not shrink the subpopulation of $H_{initial}$ by much – this could possibly be an uninteresting mined hypothesis. In contrast, if $\frac{n'}{n}$ is close to 0, then the mined hypothesis may too small and specific for it to have any statistical or practical significance.

In light of the above explanations to why a hypothesis cannot be objectively concluded to be (un)interesting, the various hypothesis mining metrics that Redhyte computes serves as a reference for the user to rank and compare the mined hypotheses based on different aspects of interestingness: trend changes, relative support, and manner of shrinkage. Ultimately, the user

would wish to rely on the domain knowledge, assisted by the metrics, to assess whether a given mined hypothesis is worthy of further investigations.

### 4.7 Future work

There is a multitude of ways towards which Redhyte can be improved. We list some possible aspects Redhyte could be improved upon in the following:

1. Scalability: Redhyte is currently hosted at shinyapps.io, for free. Better hardware is required for Redhyte to scale up to larger datasets, to truly address Big Data. Furthermore, R is not a language known for scalability and computing performance. Using the same concepts and algorithm, Redhyte can be rebuilt in a much faster language, such as Java or C++.
2. Use of other types of supervised learning models as context mining models: Random forest model is one of many types of supervised learning models. We chose to implement the random forest model in Redhyte for its accuracy and interpretability, amongst many of its advantages (see section 4.4.5 in Discussion), though more models could potentially be explored. It is also possible to conceptualize context mining in an ensemble learning manner, where multiple models are used to mine for context attributes, and a majority vote on the attributes is subsequently considered.
3. Pairwise or multiple context items for mined hypotheses: Using the mined context attributes, Redhyte constructs mined hypotheses by considering every possible context item, and insert each of them into the initial hypothesis. It is possible to insert more than one context item into the initial hypothesis; we could for instance, consider every possible pairwise combinations of context items. This leads to even more specific mined hypotheses, which could potentially be more interesting. In the current iteration of Redhyte, we only consider single context items for mined hypotheses,

mainly due to scalability issues: for smaller datasets, considering more than one context items leads to mined hypotheses with low supports and thus low statistical significance.

4. Improved hypothesis mining metrics: We note that the hypothesis mining metrics we have implemented in Redhyte are each concerned with specific aspects of interestingness: the difference lift, independence lift, and adjusted independence lift are meant to capture trend changes, the contribution and independence lift to capture relative support of mined hypotheses, and the adjusted independence lift to capture directed shrinkage. More work could be done to further develop such metrics.
5. Visualization of mined hypotheses: Mined hypotheses are scored and ranked using the hypothesis mining metrics in Redhyte. More work could be done to better visualize the mined hypotheses. For instance, Armstrong and Wattenberg (2014) introduced the comet chart for the visualization of Simpson's Reversals. Lehe and Powell at the Visualizing Urban Data Idealab (2014) gave a visually appealing representation of Simpson's Reversals. In Redhyte, the Simpson's Reversal is simply one of multiple possible types of mined hypotheses, and these visualizations can be used for mined hypotheses that do not display Simpson's Reversal as well.

## Conclusion

We have developed Redhyte, an interactive platform for rapid exploration of data and hypothesis testing, and presented a novel method using which data mining techniques can be used to complement statistical hypothesis testing. Redhyte allows the scientist and the data analyst to remain in the well-established framework of hypothesis testing, even when there is a large number of attributes in the dataset. The GUI was designed to allow non-users of statistical computing languages, such as R, to harness the power of R and modern statistical techniques. Owing to the modular structure of the system, it is possible to further expand Redhyte in different directions.

Redhyte was not designed to replace the conventional hypothesis testing framework – rather, the former should be used in conjunction with the latter. For instance, Redhyte may be used in a way such that the analyst can take selected mined hypotheses and put them under further statistical rigor, for evaluation or even confirmation. We believe Redhyte is a good addition to the arsenal of the scientist and the data analyst, by giving them an additional tool for the rapid exploration of data.

Redhyte can be found at https://tohweizhong.shinyapps.io/redhyte/, with the source codes housed in a GitHub repository at https://github.com/tohweizhong/redhyte.

## References


1. Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD Conference*.
2. Agresti, A. (1996). An Introduction to Categorical Data Analysis. (New York: John Wiley & Sons, Inc.).
3. Agresti, A. and Franklin, C. (2012). Statistics: The Art and Science of Learning from Data. (New York: Pearson Education, Inc.).
4. Akbani, R., Kwek, S. and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Proceedings of 15th European Conference on Machine Learning*, 39-50.
5. Armstrong, Z. and Wattenberg, M. (2014). Visualizing Statistical Mix Effects and Simpson's Paradox. *Proceedings of IEEE Information Visualization 2014*.
6. Batuwita, R., and Palade, V. (2011). Adjusted Geometric-Mean: A Novel Performance Measure for Imbalanced Bioinformatics Datasets Learning. *Journal of Bioinformatics and Computational Biology* **10**, 125003-1-125003-23.
7. Bickel, P., Hammel, E. and O'connell, J. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187**, 398-404.
8. Box, G. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318-335.
9. Breiman, L. (1996a). Bagging predictors. *Machine Learning* **24**,123-140.
10. Breiman, L. (1996b) Out-of-bag estimation, https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf
11. Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.
12. Breslow, N. and Day, N. (1980). Classical methods of analysis of grouped data. *Statistical Methods in Cancer Research* **1**, 122-159.
13. Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* **39**, 3446-3453.
14. Bryll, R. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition* **20**, 1291-1302.
15. Caruana, R., Karampatziakis, N. and Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25$^{th}$ International Conference on Machine Learning*, 96-103.

16. Caruana, R. and Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of 23rd International Conference on Machine Learning*, 161-168.
17. Chang, W. (2015). shiny: Web Application Framework for R. R package version 0.11. http://CRAN.R-project.org/package=shiny
18. Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321-357.
19. Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *Technical Report: Department of Statistics, University of California, Berkeley*.
20. Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 273-297.
21. DerSimonian, R., Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177-188.
22. Engle, F. (1984). Wald, Likelihood ratio, and Lagrange multiplier tests in econometrics. In Handbook of Econometrics, **2**, Griliches, Z. and Intriligator, M. ed. (Amsterdam: Elseoier Science Publishers BV.).
23. Fayyad, U. and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the International Joint Conference on Uncertainty in AI*, 1022-1027.
24. Fegel, W. (2007). The genetics of the Rhesus blood group system. *Blood Transfusion* **5**, 50-57.
25. Froehlich, F. and Kent, A. (1995). The Froehlich/Kent Encyclopedia of Telecommunications **9**. (London: CRC Press).
26. Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* **15,** 3133-3181.
27. Goethals, B. (2003). Survey on Frequent Pattern Mining. *University of Helsinki*, 1-43.
28. Gosset, W. (Student). (1908). The Probable Error of a Mean. *Biometrika* **6**, 1-25.
29. Gupta, D., Malviya, A. and Singh, S. (2012). Performance Analysis of Classification Tree Learning Algorithms. *International Journal of Computer Applications* **55**, 39-44.
30. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). (New York: Springer-Verlag).

31. Havlicek, L. and Peterson, N. (1974). Robustness of the t-test: A Guide for researchers on effects of violations of assumptions. *Psychological Reports* **34**, 1095-1 114.
32. He, H. and Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1263-1284.
33. Hideko, K. and Hiroaki, Y. (2012). Rapid Feature Selection Based on Random Forests for High-Dimensional Data. *Information Processing Society of Japan SIG Notes 2012-MPS-89* **3**, 1-7.
34. Jarman, K. (2013). The Art of Data Analysis: How to Answer Almost any Question Using Basic Statistics. (New York: John Wiley & Sons, Inc.).
35. Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* **30**.
36. Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. *Proceedings of the 14th International Conference on Machine Learning*, 179-186.
37. Lehe, L. and Powell, V. http://vudlab.com/simpsons/
38. Lichman, M. (2013). UCI Machine Learning Repository http://archive.ics.uci.edu/ml
39. Liu, G. et al. (2011) Towards exploratory hypothesis testing and analysis. *IEEE 27th Int. Conf. Data Eng*. 745-756.
40. Mann, H. and Whitney, D. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60.
41. Mantel, N. (1963). Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association* **58**, 690-700.
42. Motulsky, H. (2012). Intuitive biostatistics: a nonmathematical guide to statistical thinking. http://www.graphpad.com/support/faqid/1790/
43. Myles, A., Feudale, R., Liu, Y., Woody, N. and Brown, S. (2004). An introduction to decision tree modeling. *Journal of Chemometrics* **18**, 275-285 (2004).
44. Pavlides, M. and Perlman, M. (2009). How Likely Is Simpson's Paradox? *The American Statistician* **63**, 226-233.
45. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be

reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50**, 157-172.

46. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
47. Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review* **33**, 1-39.
48. Russell, S. and Norvig, P. (2009). Artificial Intelligence: A Modern Approach (3rd ed.). (New York: Pearson Education, Inc.).
49. Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507-2517.
50. Shapiro, S. and Wilk. M. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591-611.
51. Smyth, G., Yang, Y., Speed, T. (2003). Statistical Issues in cDNA Microarray Data Analysis. *Methods in Molecular Biology* **224**, 111-136.
52. Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology* **59**, 1-34.
53. Svetnik, V. et al. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947-1958.
54. Wang, S. (2003). Artificial Neural Network. *Interdisciplinary Computing in Java Programming, The Springer International Series in Engineering and Computer Science* **743**, 81-100.
55. Weiss, G. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations* **6**, 7-19.
56. Welch, B. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika* **34**, 28-35.
57. West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7**, 723-732.
58. Wong, L. (2014). Lecture notes for the course CS4220 Knowledge Discovery Methods for Bioinformatics, Unit 1: Essence of Biostatistics, at the National University of Singapore.
59. Yates, F. (1934) Contingency table involving small numbers and the $\chi 2$ test. Supplement to the *Journal of the Royal Statistical Society* **1**, 217-235.

## Appendix

### A. Difference lift and contribution

Given the initial hypothesis $H_{initial}$ in a 2x2 contingency table,

| | **Target attribute class $T_1$** | **Target attribute class $T_2$** |
|---|---|---|
| **Comparing attribute class $C_1$** | $c_{11}$ $(p_1)$ | $c_{12}$ |
| **Comparing attribute class $C_2$** | $c_{21}$ $(p_2)$ | $c_{22}$ |

where

$$p_1 = \frac{c_{11}}{c_{11} + c_{12}}, p_2 = \frac{c_{21}}{c_{21} + c_{22}} \quad (1)$$

Adding a context item $I = \{A_{ctx} = v_{ctx}\}$,

| $I = \{A_{ctx} = v_{ctx}\}$ | **Target attribute class $T'_1$** | **Target attribute class $T'_2$** |
|---|---|---|
| **Comparing attribute class $C'_1$** | $c'_{11}$ $(p'_1)$ | $c'_{12}$ |
| **Comparing attribute class $C'_2$** | $c'_{21}$ $(p'_2)$ | $c'_{22}$ |

where $P(T'_1) = P(T_1 \mid I)$ and $p'_1$ and $p'_2$ as similarly defined:

$$p'_1 = \frac{c'_{11}}{c'_{11} + c'_{12}}, p'_2 = \frac{c'_{21}}{c'_{21} + c'_{22}} \quad (2)$$

The first hypothesis mining metric, defined in Liu *et al*, is the difference lift, as follows:

$$DiffLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = \frac{p'_1 - p'_2}{p_1 - p_2} \quad (3)$$

The difference lift takes into account the change in trend in the mined hypothesis $H^*$, when a context item is added into the initial hypothesis $H_{initial}$. In particular, if

$$DiffLift < 0 \quad (4)$$

then a Simpson's Reversal occurs. Intuitively, for a mined hypothesis $H^*$, we would like to evaluate its interestingness at least by three different measures: i) whether the trend has been

reversed, ii) whether the change in trend (trend amplification or reversal) is substantial, and iii) whether the subpopulations of $H^*$ are still large enough. The difference lift is able to account for the first two measures, but not the third.

Liu *et al* defines the second hypothesis mining metric, contribution, as follows:

$$Contribution\ (I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = \frac{1}{p_1 - p_2}\left(\frac{n'_1}{n_1}(p'_1 - p_1) - \frac{n'_2}{n_2}(p'_2 - p_2)\right)$$
$$= \frac{1}{p_1 - p_2}\left(\frac{c'_{11} + c'_{12}}{c_{11} + c_{12}}(p'_1 - p_1) - \frac{c'_{21} + c'_{22}}{c_{21} + c_{22}}(p'_2 - p_2)\right) \quad (5)$$

Contribution, while considering the subpopulation sizes in $H^*$, loses the property that difference lift has in (4). For example, given

$$p_1 = 0.6, p_1 = 0.3, p'_1 = 0.7\ p'_2 = 0.5$$

$$DiffLift = \frac{0.7 - 0.5}{0.6 - 0.3} > 0$$

$$Contribution = \frac{1}{0.6 - 0.3}\left(\frac{n'_1}{n_1}(0.7 - 0.6) - \frac{n'_2}{n_2}(0.5 - 0.3)\right)$$

which may be negative, depending on $\frac{n'_1}{n_1}$ and $\frac{n'_2}{n_2}$.

## B. Derivation of independence lift

Consider $p_1$:

$$p_1 = \frac{c_{11}}{c_{11} + c_{12}} = \frac{\frac{c_{11}}{n}}{\frac{c_{11} + c_{12}}{n}}, n = \sum_i \sum_j c_{ij}$$

$$p_1 = \frac{P(T_1 \cap C_1)}{P(C_1)} = P(T_1 \mid C_1) \quad (6)$$

Likewise,

$$p_2 = P(T_1 \mid C_2), \qquad p'_1 = P(T'_1 \mid C'_1), \qquad p'_2 = P(T'_1 \mid C'_2) \quad (7)$$

By Bayes Theorem,

$$p_1 = P(T_1 \mid C_1) = P(T_1) \cdot \frac{P(C_1|T_1)}{P(C_1)} = P(T_1) \cdot i_1, \text{ where } i_1 = \frac{P(C_1|T_1)}{P(C_1)} \quad (8)$$

$$i_1 = \frac{P(C_1 \mid T_1)}{P(C_1)} = \frac{P(C_1 \cap T_1)}{P(C_1)P(T_1)} \quad (9)$$

By the definition of independence, $C_1$ and $T_1$ are independent if

$$P(C_1 \cap T_1) = P(C_1)P(T_1) \Leftrightarrow i_1 = 1$$

Therefore, $i_1$ is a measure of association/independence between $C_1$ and $T_1$. We would like to call $i_1$ as the *independence factor* of $C_1$ on $T_1$. $i_1$ can be easily computed from the contingency table of $H_{initial}$:

$$i_1 = \frac{P(C_1 \cap T_1)}{P(C_1)P(T_1)} = \frac{\frac{c_{11}}{n}}{(\frac{c_{11} + c_{12}}{n})(\frac{c_{11} + c_{21}}{n})} = \frac{\frac{c_{11}}{n}}{\frac{(c_{11} + c_{12})(c_{11} + c_{21})}{n^2}}$$

$$= n \cdot \frac{c_{11}}{(c_{11} + c_{12})(c_{11} + c_{21})} \quad (10)$$

Likewise, $i_2$, the independence factor of $C_2$ on $T_1$, and is a measure of association/independence between $C_2$ and $T_1$:

$$i_2 = \frac{P(C_2 \cap T_1)}{P(C_2)P(T_1)} = \frac{\frac{c_{21}}{n}}{(\frac{c_{21}+c_{22}}{n})(\frac{c_{21}+c_{11}}{n})} = \frac{\frac{c_{21}}{n}}{\frac{(c_{21}+c_{22})(c_{21}+c_{11})}{n^2}}$$

$$= n \cdot \frac{c_{21}}{(c_{21}+c_{22})(c_{21}+c_{11})} \quad (11)$$

The difference $i_1 - i_2$ can be considered to be a form of measure of the differential extent of association/independence that $T_1$has on $C_1$ and $C_2$. In the same manner, $i'_1, i'_2$ and $i'_1 - i'_2$ are defined accordingly.

We next define the independence lift to be

$$IndpLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = \frac{i'_1 - i'_2}{i_1 - i_2}$$

$$= \frac{n'}{n}\left(\frac{\frac{c'_{11}}{(c'_{11}+c'_{12})(c'_{11}+c'_{21})} - \frac{c'_{21}}{(c'_{21}+c'_{22})(c'_{21}+c'_{11})}}{\frac{c_{11}}{(c_{11}+c_{12})(c_{11}+c_{21})} - \frac{c_{21}}{(c_{21}+c_{22})(c_{21}+c_{11})}}\right) \quad (12)$$

The term $\frac{n'}{n}$ allows for $H^*$ to be evaluated on relative subpopulation sizes in comparison to $H_{initial}$. In addition, by (8),

$$IndpLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = \frac{i'_1 - i'_2}{i_1 - i_2} = \frac{\frac{p'_1}{P(T_1 \mid I)} - \frac{p'_2}{P(T_1 \mid I)}}{\frac{p_1}{P(T_1)} - \frac{p_2}{P(T_1)}}$$

$$= \frac{p'_1 - p'_2}{p_1 - p_2} \cdot \frac{P(T_1)}{P(T_1 \mid I)}$$

$$= DiffLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) \cdot \frac{1}{i_I}, \quad (13)$$

where

$$i_I = \frac{P(T_1 \mid I)}{P(T_1)} = \frac{P(T_1 \cap I)}{P(T_1)P(I)} = \frac{\frac{c'_{11} + c'_{21}}{n'}}{\frac{c_{11} + c_{21}}{n}} \quad (14)$$

(12) and (13) implies that the independence lift acquires the property of the difference lift shown in (4), while being able to account for changes in subpopulation sizes.

### C. Derivation of adjusted independence lift

Consider the following: $i_I$ is the independence factor of the context item $I$ on $T_1$. If $i_I = 1$, $T_1$ and $I$ are independent. This implies that the removal of subjects from the subpopulations of $H_{initial}$, by adding the context item $I = \{A_{ctx} = v_{ctx}\}$, to form $H^*$ is more "haphazard", as compared to if $i_I \neq 1$. We call the case when $i_I = 1$ or when $i_I$ is close to 1, *undirected shrinkage* of subpopulations. When $i_I$ is far from 1, we call that *directed shrinkage*. Under undirected shrinkage, the removal of subjects from $H_{initial}$ was not influenced by the context item $I = \{A_{ctx} = v_{ctx}\}$, and hence we might say that if $H_{initial}$ was (in)significant in the first place, then $H^*$ is likely to be (in)significant as well. Therefore, a mined hypothesis would be more interesting if $i_I$, or equivalently, $\frac{1}{i_I}$, deviates as far away from 1 as possible, i.e. direct shrinkage. Based on this intuition, we define the adjusted independence lift:

$$AdjustedIndpLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) = DiffLift(I = \{A_{ctx} = v_{ctx}\} \mid H_{initial}) \cdot \left|1 - \frac{1}{i_I}\right| \quad (15)$$

In all, Redhyte uses the above 4 hypothesis mining metrics, which are the difference lift, contribution, independence lift, and adjusted independence lift, in addition to the $\chi 2$ test statistic and the (adjusted) p-values, to evaluate the interestingness of mined hypotheses.