Redhyte: Towards a Self-diagnosing, Self-correcting, and Helpful Analytic Platform

Wei Zhong Toh, Kwok Pui Choi, and Limsoon Wong

Department of Computer Science &
Department of Statistics and Applied Probability
National University of Singapore
13 Computing Drive, Singapore 117417
tohweizhong@u.nus.edu, stackp@nus.edu.sg, wongls@comp.nus.edu.sg

Abstract. We present a platform named Redhyte, short for an interactive platform for "Rapid exploration of data and hypothesis testing". Redhyte aims to augment the conventional statistical hypothesis testing framework with data-mining techniques in a bid for more wholesome and efficient hypothesis testing. The platform is self-diagnosing (it can detect whether the user is doing a valid statistical test), self-correcting (it can propose and make corrections to the user's statistical test), and helpful (it can search for promising or interesting hypotheses related to the initial user-specified hypothesis). In Redhyte, hypothesis mining consists of several steps: context mining, mined-hypothesis formulation, mined-hypothesis scoring on interestingness, and statistical adjustments. To capture and evaluate specific aspects of interestingness, we developed and implemented various hypothesis-mining metrics. Redhyte is an R shiny web application and can be found online at `https://tohweizhong.shinyapps.io/redhyte`, and the source codes are housed in a GitHub repository at `https://github.com/tohweizhong/redhyte`.

Keywords: Statistical hypothesis testing, hypothesis analysis, hypothesis mining, data mining

1 Introduction

Much data is collected today for a variety of initial purposes. In the hands of a careful professionally-trained statistician or analyst who has a deep knowledge of the problem domain, many insights can be reliably gained from such data. However, it is often the case that an analysis project has to be carried out by someone who may lack domain knowledge or lack training, and sometimes even a professional analyst may be overwhelmed (e.g., by the volume and complexity of the data or the pressure of time) and may make mistakes [8]. A self-diagnosing, self-correcting and helpful analytic system is envisioned here to make analysis of data not only easy but also rigorous in such situations.

Self-diagnosing. All statistical tests have assumptions (e.g., observations are independent and identically distributed, observations are normally distributed)

and their conclusions are correct only when those assumptions are met. In traditional studies (e.g., a case-cohort study), subjects are carefully selected and experiments are designed so that such assumptions are met. But in today's big-data setting, we often just assemble and pull in all relevant data we could get our hands on, and there would typically be no careful selection to ensure such assumptions are met. A self-diagnosing analytic system, while making it convenient for a user to express and do a statistical test, also checks whether the test the user is doing is valid on his data. The challenging research questions include: (i) There may be no known way to check some assumptions, and thus deep statistical research is needed to figure out how to check them; (ii) the only known ways to check some assumptions are computationally costly, and thus deep algorithmic research is needed to figure out how to make these checks computationally more feasible; and (iii) how to explain to the user in a way that he can understand exactly why his requested statistical test is invalid.

Self-correcting. It is not sufficient to simply tell the user that he is performing a statistical test that is invalid on his data. The user may not know what action to take to deal with it. A self-correcting analytic system goes one step further, and tells the user how to deal with this. The challenges include: (i) How to identify alternative tests or correction steps, (ii) how to decide which alternative or correction is the most suitable one, (iii) how to explain such correction steps to the user in a way he could understand, and (iv) how to make it convenient for the user to choose and execute these corrections. Moreover, (v) for some situations, there is no known way to work around the problem, and novel idea is needed to develop the alternatives that can work in such situations.

Helpful. Beyond self-diagnosing and self-correcting, a good analytic system should also be helpful in the following sense. Initially, the user specifies a hypothesis that he wants to test. Given this initial hypothesis, the system now has some idea about what the user is interested in, and it should suggest some useful related hypotheses to the user that may give him some deeper insight into his problem. The challenges include: (i) How to identify related hypothesis, (ii) how to rank them, and (iii) how best to communicate them to the user.

In this manuscript, we describe Redhyte, which is an interactive platform for "Rapid exploration of data and hypothesis testing". Redhyte works by allowing the user to specify an initial hypothesis to be tested using one of the classical statistical tests (viz. t-test or χ^2 test), checks the validity of the test, makes corrections to the test if the initial test is detected to be invalid, as well as suggests informative related hypotheses. We believe Redhyte is a first, albeit small, step toward a self-diagnosing, self-correcting and helpful analytic platform.

The main part of this paper is organized as follows. Section 2 is a description of the Redhyte system, in particular its key functionalities. Section 3 is a case study to illustrate the features of Redhyte. The case study is based on the adult dataset from the UCI machine learning repository. Finally, Section 4 summarizes the work and discusses related works.

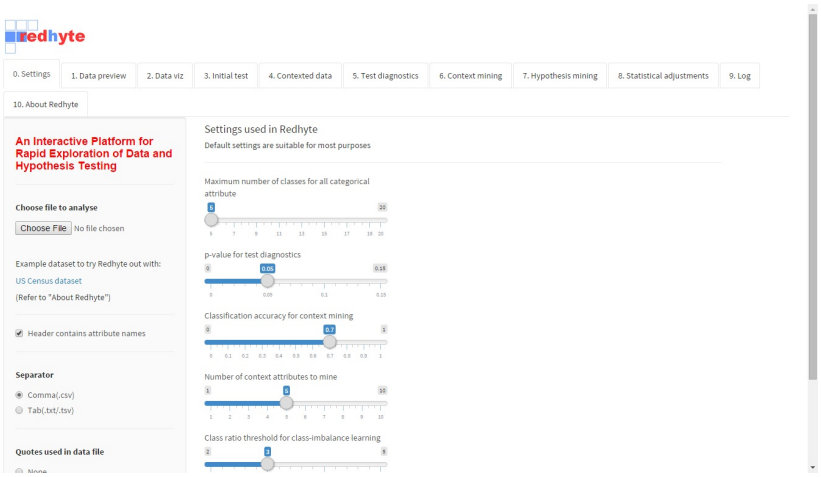

Fig. 1. A screenshot of Redhyte.

2 System Description

This section describes various modules that are more vital in Redhyte's workflow (and user-friendly interface); the less vital modules are omitted.

2.1 User Interface

Redhyte is fundamentally a web application that renders in a web browser, such as Google Chrome or Mozilla Firefox. Redhyte's user-facing interface is organised into tabs, as shown in Figure 1, with each tab housing a specific functionality that Redhyte provides. The tabs are ordered from left to right, mirroring the expected workflow of an analysis: the user makes some brief checking and exploration of his input data (the data-preview and data-visualization modules), the user specifies his initial hypothesis (the initial-tests module), the user looks at the validity assessments of the test on his hypothesis (the test-diagnostic module), the user gets information on factors that could strengthen or contradict his hypothesis (the context-mining module), and the user looks at related hypothesis mined by Redhyte (the mined-hypothesis formulation and scoring module).

2.2 Initial Hypothesis Set-Up and Tests

After loading the input dataset into the platform, the user is prompted to set up the initial hypothesis. To establish a common lingo between the user and Redhyte regarding the initial hypothesis and all other steps in the Redhyte workflow, Redhyte defines the following terms in a hypothesis: "target attribute", "comparing attribute", and "context attribute". As an example, a hypothesis comparing resting heart rate between smokers and non-smokers amongst the

males only, would have resting heart rate as the target attribute, smoking status as the comparing attribute, and gender as the context attribute. In addition, an {attribute = value} pair, e.g. {gender = male}, is called a context item. After the hypothesis is set up, depending on whether the target attribute is numerical or categorical, an initial t-test or χ^2 test is used to assess the hypothesis. Naturally, the user uses the Redhyte graphics interface to specify his hypothesis.

Following Liu et al. [11], we write $H = \langle P, A_{diff} = v_1|v_2, A_{target} = v_{target}\rangle$ to denote a hypothesis H. The set of items P is the context, which is the subset of the subjects in the dataset satisfying all items in P. The comparing attribute is A_{diff}, and $P_1 = P \cup \{A_{diff} = v_1\}$ and $P_2 = P \cup \{A_{diff} = v_2\}$ define the two subpopulations to be compared. The target attribute, on which the two subpopulations is being compared, is A_{target}.

2.3 Test Diagnostics

The test diagnostics tab aims to work on several issues:

- If the initial test is a t-test [7], the assumptions of normal distributions and equal variances are checked, using the Shapiro-Wilk test [16] and F-test [5] respectively. If either test is significant, the initial test is re-assessed using the Wilcoxon rank sum test [12], the non-parametric equivalent of the t-test.
- If the initial test was a "collapsed" χ^2 test [14], Redhyte computes the individual χ^2 contributions of each class in the comparing attribute. A collapsed χ^2 test refers to a χ^2 test where one or both groups in the initial hypothesis consist of more than one class of the comparing attribute.
- In both cases, the Cochran-Mantel-Haenszel test [3] is further used on other attributes in the dataset, to assess whether the effect of the comparing attributes on the target attribute is influenced by these co-variates.

2.4 Context Mining

In order to identify potential confounding attributes to the initial hypothesis, Redhyte uses classification models. Specifically, it constructs two random-forest [2] models to predict, using all other attributes not involved in the initial hypothesis as predictors or covariates, the target and comparing attributes. The idea is: If an attribute A is able to help classify the target or the comparing attribute, then A might possibly be related to either attribute. Thus it might be interesting to consider using A as a context attribute for the initial hypothesis. The random-forest models confer a measure of variable importance, ranking the attributes according to how well they contribute to the classification of the target and comparing attribute. Shortlisting the top few attributes from the variable importance measure gives the mined context attributes.

2.5 Mined-Hypothesis Formulation and Scoring

After shortlisting the mined context attributes, each attribute is used as a context attribute and inserted into the initial hypothesis to form mined hypotheses, by means of stratification. For example, if occupation is a mined context

attribute, then examples of mined hypotheses could be restricting the initial hypothesis to all teachers only or all engineers only.

Stratification due to a mined hypothesis can bring about one of three outcomes: The trend observed in the hypothesis could either be (i) amplified, (ii) unchanged, or (iii) reversed (Simpson's reversals). After forming the mined hypotheses, they are ranked using the various hypothesis-mining scores (difference lift, contribution, independence lift, and adjusted independence lift) to evaluate which of these three outcomes they fit. The "difference lift" and "contribution" scores are given by Liu et al. [11]. The former compares a hypothesis $H = \langle P, A_{diff} = v_1 | v_2, A_{target} = v_{target} \rangle$ with a new hypothesis $H^* = \langle P \cup \{A = v\}, A_{diff} = v_1 | v_2, A_{target} = v_{target} \rangle$, which has an extra item $\{A = v\}$ in its context, to see whether the trend specified in H has changed (amplification or reversal) substantially in H^*. The latter measures the change in trend in a way that is weighted by the subpopulations being compared in H and H^*. We find that there are situations where this score disagrees completely with difference lift (e.g., the former sometimes reports a negative change in trend when the latter reports a positive change). So we also use the "independence lift" and "adjusted independence lift" scores, which are defined in the full Redhyte report [18]. These two scores always agree with difference lift in the direction of the change in trend while also take into consideration the sizes of the subpopulations being compared in H and H^*. Due to space contraint, definitions and detailed treatment [18] of these scores are omitted here.

2.6 Statistical Adjustments

Besides inserting mined context items into the initial hypothesis to detect issues like Simpson's reversals, these mined context attributes can also be held accounted for using regression models. The regression model is constructed using the target attribute as the response variable and the mined context attributes as predictors. We call this resultant model the adjustment model. For numerical target attributes, linear regression is used as the adjustment model [6]. For categorical target attributes, the logistic regression model is used instead [4].

To construct the adjustment model, Redhyte first uses the stepwise regression algorithm to further shortlist, from the mined context attributes, a subset of them to be used for adjustments in the adjustment model. Next, the construction of the adjustment model and its use are as follows:

- For numerical target attributes, the target attribute is used as the dependent variable and the shortlisted mined context attributes, with all pairwise interaction terms, are used as predictors / covariates. The constructed adjustment model gives the required numerical adjustments of the target attribute (computed as actual values found in dataset minus fitted values from model). A t-test is then done on the numerical adjustments, to compare with the initial t-test.
- For categorical target attributes, the target attribute is used as the dependent variable, while the shortlisted mined context attributes and the comparing attribute, with all pairwise interaction terms, are used as predictors /

covariates. The constructed adjustment model lends itself a way to conduct "what-if" analysis (i.e., what if the entire dataset consists of samples that differ only in the target and the comparing attribute?) For instance, if the mined context attributes are gender and occupation, we ask: What if the entire dataset consists of samples that are all males and all engineers? The logistic regression model provides a means for such an analysis, by "substituting" these covariate values into the model equation.

3 Use-Case

In this section, we use the adult dataset (from the UCI machine learning repository, `http://archive.ics.uci.edu/ml`) to illustrate a simple use-case for hypothesis mining by Redhyte. The adult dataset contains the demographical data of 32,561 adults. The target attribute in this dataset is the binary income attribute, taking these values: $> 50K$ and $\leq 50K$. Consider the hypothesis below:

> 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} and {Craft-repair}?

3.1 Initial Test

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 Figure 2.

	Income >50K	Income ≤50K	Total
Adm-clerical	439 (14.2%)	2645 (85.8%)	3084
Craft-repair	844 (22.8%)	2850 (77.2%)	3694
Total	1283	5495	6778

Fig. 2. Contingency table of the initial hypothesis.

Using 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 contingency tables in Figure 3, which illustrate two instances of a Simpson's Paradox [17], with both genders and workclass resulting in reversals of the trend in Figure 2.

3.2 Hypothesis-Mining Metrics

The hypothesis-mining metrics evaluated on the three items ({Sex = Male}, {Sex = Female}, and {Workclass = Self-emp-not-inc}) are given in Figure 4.

{Sex = Male}	Income>50K	Income≤50K	Total
Adm-clerical	251 (24.2%)	787 (75.8%)	1038
Craft-repair	829 (23.5%)	2695 (76.5%)	3524
Total	1080	3482	4562

{Sex = Female}	Income>50K	Income≤50K	Total
Adm-clerical	188 (9.2%)	1858 (90.8%)	2046
Craft-repair	15 (8.8%)	155 (91.2%)	170
Total	203	2013	2216

{Workclass = Self-emp-not-inc}	Income>50K	Income≤50K	Total
Adm-clerical	16 (34.8%)	30 (65.2%)	46
Craft-repair	90 (18.0%)	409 (82.0%)	499
Total	106	439	545

Fig. 3. Contingency tables of mined hypothesis with {Sex = Female}, {Sex = Male}, and {Workclass = Self-emp-not-inc}.

Context items	Diff lift	Contrib	Indep lift	Adjusted indep 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

Fig. 4. Hypothesis-mining metrics evaluated for the selected context items.

Based on our initial hypothesis, 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.

3.3 Statistical Adjustments

Following through with statistical adjustments in Redhyte, sex, relationship, workclass, and education are recommended for use in statistical adjustment. Since the target attribute, income, is a categorical one, the adjustment model is a logistic regression model, to be used for "what-if" analysis. In particular, as shown in Figure 5, after adjusting for {sex = Male}, {relationship = Husband}, {workclass = Self-emp-not-inc} and {education = Bachelors}, a Simpson's reversal is observed in the hypothesis that administrative clerks earn more than craft repairers.

Fig. 5. A visualization show the proportions of administrative clerks and craft repairers earning more than 50k, before (left chart) and after (right chart) adjusting for sex, relationship, worksclass, and education.

4 Conclusion

Hypothesis testing is one of the mainstay tools in data analysis, as it allows the analyst to make comparisons between different groups of samples. Conventionally, data-analysis workflows primarily consist of the following steps: (i) have a scientific question in mind, (ii) formulate an assertion or hypothesis, (iii) collect and clean relevant data, and finally (iv) test the hypothesis using statistical techniques, in order to decide whether to reject the hypothesis. Putting together a hypothesis with a statistical test allows the analyst to make justifiable conclusions from the data, and this process is often prompted by the initial question or hypothesis in mind. In other words, collection of data in conventional data analysis settings are often driven by domain requirements and scientific questions a priori.

From a statistical viewpoint, having some initial scientific questions to drive the collection of data confers an important upshot: The collected data is well specified. To be more precise, with proper sampling methods, issues such as lack of independence, dissimilar distributions, unequal variances, class imbalance etc. can be addressed and alleviated. However, the big-data setting brings about two interesting scenarios, specifically the collection of data without a scientific question a priori, and the "large p, small n" phenomenon [20].

Data collected without any initial scientific questions poses a problem: Assumptions of many statistical techniques, including hypothesis testing, are more often violated than not. Moreover, having a large number of attributes in a dataset requires adequate treatment and analysis to properly account for these attributes. 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 confounding factors). 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 Simpsons Reversal [13]. As in the use case presented earlier, while the initial observation suggests the hypothesis that craft repairers earn more than administrative clerks (cf. Figure 2), after automatically detecting and adjusting for confounding fac-

tors by Redhyte, the completely opposite hypothesis that administrative clerks earn more than craft repairers emerges (cf. Figure 5). 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 [1].

In this paper we have introduced Redhyte, a platform for statistical hypothesis testing on datasets collected without initial scientific questions. The workflow in Redhyte is as follows: (i) User first suggests an initial hypothesis, which could be rough, intuitive, and domain knowledge-driven. (ii) Redhyte first works on some initial statistical test on the initial hypothesis, and assesses the validity of the statistical test applied to the initial hypothesis. (iii) Redhyte uses data-mining techniques to search for potential confounding attributes (context mining), and uses them to form variants of the initial hypothesis, by means of stratification. These variants of the initial hypothesis are then scored and ranked, to let the user home in on the more interesting ones. (iv) Finally, Redhyte attempts to adjust for these potential confounding attributes using regression techniques.

It is easy to make mistakes involving statistics. Powerful statistical tools—R, Minitab, SPSS, etc.—certainly remove a lot of the difficulty of doing statistical calculations. However, they do not check whether the user is applying the statistical tests correctly. The special aspects of Redhyte compared to commonly-used statistical tools are that it explicitly supports checking whether what the user is doing is valid, guiding him to do his statistical test correctly, and recommending to him related hypotheses that might lead to some deeper insight. Redhyte is thus a step towards building a self-diagnosing, self-correcting, and helpful analytic system, albeit it current supports only very simple statistical tests.

On the data-mining side, the closest research related to Redhyte is perhaps the work of Liu et al. on exploratory hypothesis testing and analysis [9, 11]. In these works, algorithms for mining and visualization of hypotheses from large datasets are described. More importantly, they also presented algorithms for linking together related hypotheses and measures for ranking hypotheses (and we have proposed refinements of these [18] and implemented them in Redhyte). Works from the data-mining field on the clustering and grouping of frequent itemsets and association rules [21, 19, 15, 10] may also be useful for generating related hypotheses (we do not consider these clustering methods here because we start from a single user-specified hypothesis, and so face a much lower mining and clustering complexity than these methods). Moreover, these methods are not concerned with ensuring the validity of a user's statistical test or guiding him toward a valid statistical test.

Acknowledgments. This work was supported in part by a Singapore Ministry of Education tier-2 grant (MOE2012-T2-1-061).

References

1. Bickel, P. Hammel, E., O'connell, J.: Sex bias in graduate admissions: Data from Berkeley. Science 187, 398–404 (1975)

2. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
3. Cochran, W. G.: Some methods for strengthening the common χ^2 tests. Biometrics 10, 417–451 (1954)
4. Cox, D. R.: The regression analysis of binary sequences (with discussion). Journal of Royal Statistical Society B 20, 215–242 (1958)
5. Fisher, R. A.: On a distribution yielding the error functions of several well-known statistics. Proceedings of the International Congress of Mathematics 2, 805–813 (1924)
6. Freedman, D. A.: Statistical Models: Theory and Practice. Cambridge University Press (2009)
7. Gosset, W. S.: The probable error of a mean. Biometrika 6, 1–25 (1908)
8. Ioannidis, J. P. A.: Why most published research findings are false. PLoS Medicine 2, e124 (2005)
9. Liu, G., Suchitra, A., Zhang, H., Feng, M., Ng, S. K., Wong, L.: AssocExplorer: An association rule visualization system for exploratory data analysis. In: Proceedings of 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1536–1539 (2012)
10. Liu, G., Zhang, H., Wong, L.: A flexible approach to finding representative pattern sets. IEEE Transactions on Knowledge and Data Engineering 26, 1562–1574 (2014)
11. Liu, G., Zhang, H., Feng, M., Wong, L., Ng, S. K.: Supporting exploratory hypothesis testing and analysis. ACM Transactions on Knowledge Discovery from Data 9, article 31 (2015)
12. Mann, H. B., Whitney, D. R: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18, 50–60 (1947)
13. Pavlides, M., Perlman, M.: How likely is Simpson's paradox? The American Statistician 63, 226–233 (2009)
14. Pearson, K.: 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 Series 5 50, 157–175 (1900)
15. Poernomo, A. K., Gopalkrishnan, V.: CP-summary: A concise representation for browsing frequent itemsets. In: Proceedings of 12th ACM SIGKDD International Conference on Knowlegde Discovery and Data Mining, pp. 687–696 (2009)
16. Shapiro, S. S., Wilk, M. B.: An analysis of variance test for normality (complete samples). Biometrika 52, 591–611 (1965)
17. Simpson, E. H.: The interpretation of interaction in contingency tables. Journal of Royal Statistical Society B 13, 238–241 (1951)
18. Toh, W. Z.: Redhyte: an interactuve platform for rapid exploration of data and hypothesis testing. Project report, National University of Singapore (2015) `http://www.comp.nus.edu.sg/~wongls/psZ/tohweizhong-fyp2015.pdf`
19. Wang, C., Parthasarathy, S.: Summarizing itemset patterns using probabilistic models. In: Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 730–735 (2006)
20. West, M.: Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statistics 7, 723–732 (2003)
21. Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing itemset patterns: A profile-based approach. In: Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 314–323 (2005)