Requirements of Statistics in Medicine

Short Title Character Count: 56/70 characters

Manuscript Word Count: 2062/3000 words

Abstract Word Count: 162/250 words

Keywords (MeSH) Up to six: Crowdsourcing, Epidemiology, Machine Learning; Unsupervised Learning; Supervised Learning; Support Vector Machine

Reference Style:
3. Oestreicher, N. The cost of adjuvant chemotherapy in patients with early-stage breast carcinoma, Cancer 2005; 104(10), pp. 2054 - 2062, DOI: 10.1002/cncr.21464.

Figure Format: Figures should be uploaded as separate figure files, in .tiff and 800dpi

Manuscript submission format: .doc

Data Repository: sr2c.case.edu

Full Title: Crowdsourcing in Epidemiology: TextM,

Short Title: Analytics for Crowdsourcing: Application to Epidemiology

Commented [KC1]: It appears that something is missing here.

ABSTRACT

Crowdsourcing has become a valuable tool to amass vast amounts of data rapidly for scientific research. This research was in response to an invited solicitation to investigate *Trends of Crowdsourcing in Epidemiology (CrowdEpi)*. To obtain data on CrowdEpi, we developed a Java-based text mining machine ``TextM" to collect relevant information from the articles queried by an XML crawler from five online sources. To study the trends, we identified four utilities within the CrowdEpi, which could be used for both subgroup and full group analyses. To automate classification for future applications, we investigated the best classifier for automatically grouping articles into these four CrowdEpi areas and the fifth, nonCrowdEpi area. However, due to non-standardization for crowdsourcing terminology and approaches, multiple utilities may be used in one article and the best classifier that forces a single-group crisp decision would have an unacceptable prediction error rate. We, therefore, devised a new intelligent, adaptive classifier iDKR which classifies articles into Decision (one of the utilities), Refinement (some utilities), and Rejection (unclear), to allow further evaluation of these articles to reach an excellent classification rate. In conclusion, we identified challenges for Crowdsourcing in Epidemiology, provided vehicles for studying CrowdEpi. Our TextM is of interest itself. TextM can serve as a general-purpose text mining tool that is customizable to a particular study, either for future trends of CrowdEpi, or a different area of application. Our iDKR calls for a paradigm change in classification, with excellent potential for decision support in medical science. Medical errors are costly, and it is common to perform further tests, surgery or a pathology study to determine a medical diagnosis. This work is a product of interfacing statistics and computer science for data science.

Commented [KC2]: This is never mentioned or described in the text.

Keywords: Knowledge Discovery, Development, Text Mining Machine, Crowdsourcing

1 INTRODUCTION

Crowdsourcing is an increasingly popular and important research tool in scientific inquiries. Crowdsourcing outsources tasks that either require input from the masses, or seek intelligence from the general public. In epidemiology, crowdsourcing is used in a variety of applications from medical imaging to health surveillance [1,2]. Crowdsourcing in epidemiology provides not only a faster and cheaper alternative to traditional methods of collecting data but also offers a means for completing tasks that were originally not possible with limited resources [3]. While the wide variety of crowdsourcing applications are exciting, crowdsourcing has unique issues. First, finding a representative sample of the target population can be difficult for crowdsourced surveys [4]. Second, tasks completed online need to be very carefully designed and moderated to minimize selection and other biases [5]. Third, crowdsourcing is typically considered a tool for conducting research instead of the main research focus, therefore key components of quality control are not always well thought out or executed. Fourth, the explosion in popularity of crowdsourcing has led to a plethora of inconsistent crowdsourcing terminology that makes uncovering relevant trends difficult. Here, we (1) find relevant data with strategic queries, using an XML-crawler and developing our own java-based parsing program, (2) propose utilities of crowdsourcing in epidemiology to both inform the breadth of research and classify the research in this field for their subgroup studies and quality control recommendations, (3) evaluate trends of crowdsourcing in epidemiology, and (4) address the requisite quality control methods used with this technique. We demonstrate our java tool, which can be reused to assess utilities in the literature, as well as facilitate better understanding of the trends that are common in epidemiological crowdsourcing. Our work also is a start towards standardizing crowdsourcing in epidemiology.

2 METHODS

In this section, we first clarify our crowdsourcing definitions and describe our methods for completing our study objectives. The overall scheme of the study is shown in Fig 1.

Commented [KC3]: I didn’t find a Fig. 1.

2.1 Crowdsourcing Definitions

We define crowdsourcing as gathering information generated by or obtained from online users to complete a desired task. The users can either knowingly contribute content to the task or passively post data via a forum or a social media website. This definition is in contrast to prior definitions of crowdsourcing, which did not include user generated content from the internet [6,7]. However, mining online user-generated data has become an important new topic and a rich source of data in crowdsourcing [8,9]. Thus, data mining passively posted content from the internet is now required for inclusion in the definition of crowdsourcing.

A crowdsourcing site is defined as a website specifically created for crowdsourcing purposes. Examples of this include Amazon Mechanical Turk [10], Foldit [11], and PatientsLikeMe [12].

2.2 Information Retrieval: Identification of Data Sources

For this study, we included articles that were peer reviewed and identified to be at the intersection of epidemiology and crowdsourcing (Figure 1). To obtain as complete a representation of epidemiological crowdsourcing articles as possible, we searched five databases: Medline, Google Scholar, Pubmed (abstracts and keywords only), Pubmed Central (full text) and MathSciNet. The query terms "(crowdsourcing OR crowdsource OR crowdsourced OR macrotask OR microtask) AND (disease OR health OR epidemiology)" were used to retrieve relevant data sources, and query terms "(crowdsourcing OR crowdsource OR crowdsourced OR macrotask OR microtask)" were used to retrieve non-crowdsourcing/epidemiology sources for training purposes.

Commented [KC4]: Missing

Commented [KC5]: You should define what you mean by this. Is it only epidemiology that is non-crowdsourcing, or is it crowdsourcing that is not epidemiology?

2.3 Information Retrieval: Building a Corpus

We developed an extensible markup language (XML) crawler to retrieve targeted articles from Medline. It was guided by a rule-based configuration that identified articles by both their mesh term and the desired crawling depth. These rules performed an initial screening of the articles in order to best identify the appropriate documents for the final corpus. The crawler started the navigation by retrieving the MESH term "Crowdsourcing" using rentrez() from the XML library in R. This program harnesses NCBI's EUtils API for parsing databases such as GenBank and PubMed [13,14]. Based on the matching mesh term, the crawler extracted the data based on the nested XML metadata from PubMed and Pubmed Central of PMID, Journal, Publication Date, Article Year, Article Title, and Author Name (First, Last). The extracted data were subsequently processed by plyr() from the R library and automatically compiled into a csv file [15]. For the other databases (e.g. Medline, Google Scholar,

Commented [KC6]: This is not clear. I think it should be split into 2 sentences. After Pubmed Central, start a new sentence: The data extracted were ….

MathSciNet), the articles were manually identified via the aforementioned query terms. Every article identified from these queries was manually downloaded and automatically converted to a text file using a Ruby program [16]. Articles that could not have their full text retrieved were removed from the final database of articles, as well as any duplicate articles.

2.4 Utility Definitions

Through a combination of data search, literature review, and domain expertise, our proposed utilities of epidemiological crowdsourcing are:

Commented [KC7]: 1, 2, and 3 look like uses, but 4 looks like a concern. If I understand correctly, 1, 2, and 3 are reasons researchers use crowdsourcing. Item 4 does not seem to fit.

Utility 1: Completing non-personal tasks. Participants are asked to complete a task that the requesters, the creators of the task, could theoretically do themselves given enough time and the appropriate skill set. The task content is not related to a participant directly. The may be simple to execute, or require a certain level skill in a particular domain.

Utility 2: Data mining independently posted content. Data is collected from information posted by users as part of a text conversation, a post, or self-entered personal data. This data can be posted on social media sites, such as Twitter or Facebook, or crowdsourcing websites, such as PatientsLikeMe and 23AndMe, or on general websites such as Github or Foursquare.

Utility 3: Providing personal information. People are recruited online and asked for information related to their opinions, knowledge, demographic information, or other relevant information. Participants are aware of that they are submitting this information and this task is typically conducted in a survey format. This task requires the input of a representative sample of participants.

Utility 4: Addressing quality control. The study specifically addresses quality control methods related to their crowdsourcing task. The quality control used in the study is reasonably well thought out and implemented.

2.5 Information Extraction: Dictionary and Java Program Development

We identified journal articles in the field of epidemiology and crowdsourcing that belonged to each utility by constructing a dictionary of relevant terms that distinguished each utility (Figure 1). The dictionary was developed to implement two stage filtering. The first stage was to screen for articles that did and did not meet the inclusion criteria of epidemiology and crowdsourcing. These articles were further classified by each of the four utilities. If the number of occurrences in a category was four or larger, the article was considered to be included in that category.

Commented [KC8]: Missing

To classify the articles we developed the Java-based software tool, Utilitary, which allows a user to input their own custom dictionaries. The tool can parse a corpus based on an imported dictionary and researcher-defined classification scheme, and output the word frequency or number of words in a category that appeared in a journal article in a .csv file format.

2.6 Training Dataset Development

For our training dataset, one-third (N=63) of the articles were selected from the final corpus to be manually annotated by 3 independent raters. Each article was assigned a label as either Utility 1, 2, 3, 4, or “Bogus”, where a “bogus” article indicated that the article did not fall under the required intersection of epidemiology and crowdsourcing. After calculating inter-annotator agreement, all cases in which disagreement was present were reviewed collaboratively by the 3 annotators and consensus was reached by a final decision moderated by the principal investigator. In addition, one of the annotators reviewed the validation sample and graded them using the same classifications described above.

2.7 Modeling and Analysis

We initialized our models with a linear SVM trained with the sample of labeled articles from the full corpus. We used weights to delineate between the hard boundaries of “Bogus” versus “Crowdsourcing/Epidemiology’ articles, which was an aggregate label applied to the articles identified as Utility 1 through 4. As a separate test, we used linear discriminant analysis, a method that finds features to differentiate between different classes of data [17]. We compared results with quadratic discriminant analysis (QDA), which relaxes some of the assumptions of LDA, and determined that LDA was more appropriate for identification of the hard boundaries of “bogus” versus

"Crowdsourcing/Epidemiology" articles.

We assessed the accuracy of the SVM and LDA classification outcomes using both 2-fold and 10-fold cross validation repeated 10 times [18,19]. An acceptable misclassification rate varies depending on application, but here we used a 20% misclassification rate as our standard for sufficient accuracy.

Finally, we corroborated our findings with exploratory projection pursuit [20]. Projection pursuit is a non-parametric tool that helps to explore interesting nonlinear structures such as clusters and separations of a high-dimensional data set by projecting the data onto a low-dimensional space.

Projection pursuit was through GGobi software system, all other data analyses were performed with R 3.1.2 [13,21].We used the SVM version and the LDA version implemented in the e1071() and MASS() R package libraries, respectively [22,23].

3 RESULTS

3.1 Descriptive Statistics of Data by Utilities

Outcomes from the automatic classification Java program indicate that using "Crowdsourcing" and "Epidemiology" as initial screening for relevant literature is appropriate, but required additional fine tuning as up to 20% (38/187) of the articles were identified by the Java Program to be "bogus" given that they did not meet either the initial screening criteria of "Crowdsourcing/Epidemiology" or did not meet any of the four utility definitions. The field was best represented by Utility 3 (33/187) where 18% of the literature was dedicated to crowdsourcing information and knowledge from a targeted, relevant population. The least represented utility was, as expected, crowdsourcing with social media (27/187) at 14% of our sampled articles.

Commented [KC9]: These numbers don't make sense. If there are five categories (Bogus and 4 utiilities), then an equal division would put 37-38 in each category, but you say the best had only 33 in it. Further, that isn't much different from the least (27).

3.2 Accuracy of Classification Models

We used a linear kernel and grid search to find optimal values of the kernel parameter, and the cost parameter C, which control the tradeoff between false positives and false negatives. The parameters chosen that optimized the accuracy of the SVM on the training set under 10-fold and 2-fold cross validation. Model training required approximately 30 minutes to complete the entire grid search. After

Commented [KC10]: This is not a sentence.

the optimal SVM parameters were found, we applied it to our test sample to yield probability estimates for appropriate classification of “Bogus” versus “Crowdsourcing/Epidemiology” papers and compared our estimates to our gold-standard manual annotations. The best model used cost parameter γ = 0.01 and C=0.7 with informative prior class weights set at 0.476 and 0.524 for “Bogus” and “Crowdsourcing/Epidemiology” articles, respectively. Our fitted misclassification rate with the linear kernel was 7%, but when predicted on new data the misclassification rate increased to 25%. Similarly, our LDA classifier also reported a fitted misclassification rate at 7%, but when predicted on new data the misclassification rate increased to 50%. The LDA classifier reported the best discriminatory split, and this decision boundary was confirmed by projection pursuit. Although the classification rate among the fitted models was excellent, predicted rates among the test set were problematic, for reasons we will outline below.

Commented [KC11]: Should this be kernel?

4 DISCUSSION

4.1 Overview

The field was least represented by studies that used crowdsourcing by leveraging social media. This utility has been deliberately omitted by prior studies of crowdsourcing [6,24], specifically because they did not feel that this utility met their definition of crowdsourcing at the time. Given the rapidly evolving field, this utility can definitely be used to characterize peer-reviewed health research.

4.2 Limitations

A relevant issue regarding our corpus is its heterogeneity. Specifically, we determined that articles that described articles from a broad, introductory, or general perspective bloated the false discovery rates. Furthermore, the addition of terms to our dictionary did not improve lookup nor classification rates amongst utilities. These false discovery rates, however, led us to discover a wide berth of jargon associated with crowdsourcing.

Commented [KC12]: Difficult to understand

Commented [KC13]: Amount?

In addition to heterogeneity, there was a paucity of articles to choose from that met the merged criterion of “Crowdsourcing + Epidemiology.” (Figure 3). As of March 28th, 2016 there are almost 500,000 full-text articles in the field of epidemiology and over 200 articles identified as “Crowdsourcing” available on PubMed platform [25]. Extracting from this larger corpus of 500,000 articles of epidemiological

Commented [KC14]: Not available. And there is no reference to a Figure 2.

studies would not be appropriate for training our models, as it would have biased the classification model away from our goal of identifying articles that used the methodology of “Crowdsourcing” for health-based research.

4.3 Conclusions and Recommendations

Looking forward, the different manifestations of the crowdsourcing paradigm open up many new avenues for scientific exploration. Taking advantage of the crowdsourcing technique requires careful adherence and thoughtfulness to protocol and quality control to obtain valid results. An advantage of applying a classification tool to the rapidly evolving crowdsourcing literature is to promote automatically scalable utilities. This avoids pre-emptively restricting the rapidly expanding impact of crowdsourcing to a limited set of features.

The definition of crowdsourcing continues to mature and expand in the literature, as new areas of research and applications emerge. An important question is how to address the wide variety of interpretations of crowdsourcing as the technique continues to gain exposure from scholars of various fields. To barely scratch the surface of descriptions, crowdsourcing platforms can be described as microwork or platforms or online labor markets or system. Participants in crowdsourcing can be called users, microworkers, macroworkers, or workers. Tasks are called assignments, games, or stratified into micro and macro tasks. Fortunately a similar, parallel issue transpired in the field of statistical programming with the R language. In 2012 The R Journal published a detailed exploration of the state of statistical package naming conventions, which illuminated wildly different approaches such as all lower cases, separation by a period or underscore, and various combinations of uppercase and lowercase [26]. The R foundation called for implementation of a consistent coding convention, arguing that the conventions were solely a matter of taste and habit and without a well-defined consistency the lack of consensus would continue. This implementation was coupled with a major push in the scientific community at the time for establishing necessary and sufficient conditions for reproducible research to encourage cumulative knowledge development [27].

Commented [KC15]: I don’t understand this.

In the spirit of the efforts of the R community and the call for reproducible research, we seek to identify common denominator syntax to encourage efforts devoted to crowdsourcing to achieve massive scalability. This is because the broader scientific community cannot stand to gain from the important health-related research questions being answered by crowdsourcing if the articles are published without

Commented [KC16]: Will not benefit?

a governing standard. This will continue to be a problem unless there is a unified definition and dictionary for crowdsourcing proposed by the American Mathematics Association or the Association for Computing Machinery. Like with programming, it is important to remember "good style" with crowdsourcing. The reason for this being though publications in the field have limited number of authors, these publications will usually have a wide number of readers. Therefore it is a good idea to agree to a common "style" up front, as proposed below:

Commented [KC17]: Like the ones proposed …?

Commented [KC18]: These two sentences are poorly worded.

Table 1: Crowdsourcing in Epidemiology

Commented [KC19]: The table needs to be explained. And it is incomplete.

	Determine Eligibility	**Recommended Syntax**	**Method Type**	**Avoid**
Study Design	Information garnered from a collection of people that were not deliberately enrolled like a clinical trial	Crowdsourcing		
Utility Type	how the data is being collected: individual, social media, cell phones	Mode of collection		
Task Type	Identify how the task is being completed or	Task		
Platform Choice	Identify if it is an established platform such as Amazon Turk, crowdflower, qualtrics, surveymoneky or gaming website	Crowdsourcing platform	Either hosted via an established platform or individually created website	Choosing hosting based on popularity in the literature, instead of on the best choice for your task type or data collection
Population		Participants		Avoid using proprietary-based language or field-specific jargon of 'workers' or 'microworkers' or 'users'
Quality Control				
Statistical Testing				

REFERENCES

1. Chunara R, Chhaya V, Bane S, Mekaru SR, Chunara R, Chhaya V, et al. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. Malar J. 2012;11(1):1–7.
2. Maier-Hein L, Mersmann S, Kondermann D, Stock, C Kenngott, H.G. Sanchez A, Wagner, M. Preukschas, A. Wekerle AL, Helfert S, et al. Crowdsourcing for reference correspondence generation in endoscopic images. Int Conf Med Image Comput Comput Interv. 2014;Springer I:349–56.
3. Daneman N, Gruneir A. Prolonged antibiotic treatment in long-term care: role of the prescriber. JAMA Intern Med. 2013;173(8):673–82.
4. Wright K. Researching Internet based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and. J Comput Commun. 2005;10(3):00–00.
5. Alonso O, Baeza-Yates R. Design and implementation of relevance assessments using crowdsourcing. Adv Inf Retr. 2011;
6. Ranard BLB, Ha YPY, Meisel ZZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med. 2014;29(1):187–203.
7. Estellés-Arolas E, González-Ladrón-de-Guevara F. Towards an integrated crowdsourcing definition. J Inf Sci. 2012;38(2):189–200.
8. Barbier G, Zafarani R, Gao H, Fung G, Liu H. Maximizing benefits from crowdsourced data. Comput Math Organ Theory. 2012;18(3):257–79.
9. Xintong G, Hongzhi W, Song Y, Hong G. Brief survey of crowdsourcing for data mining. Expert Syst Appl. 2014;
10. Amazon. Amazon Mechanical Turk [Internet]. 2015 [cited 2015 Jul 1]. Available from: https://www.mturk.com
11. Center for Game Science at University of Washington UD of B. Foldit [Internet]. 2015 [cited 2015 Jul 1]. Available from: https://fold.it/portal/
12. Patients Like Me. PatientsLikeMe [Internet]. 2015 [cited 2015 Jul 1]. Available from: https://www.patientslikeme.com/
13. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing. 2014. Available from: http://www.r-project.org/
14. Winter D. rentrez: Entrez in R [Internet]. R Package Version 0.4.1. 2015. Available from: http://cran.r-project.org/package=rentrez
15. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. J Stat Software2. 2011;40(1):1–29.
16. Elsaid E. PDF to Text converter using ruby [Internet]. 2014 [cited 2015 Jun 1]. Available from: http://www.dzone.com/snippets/pdf-text-converter-using-ruby
17. Ripley BD. Linear Discriminant Analysis. In: Pattern Recognition and Neural Networks. 1996. p. 3.
18. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Jt Conf Artif Intell. 1995;14(2):1137–45.
19. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2009.
20. Sun J. Some practical aspects of exploratory projection pursuit. SIAM J Sci Comput.

1993;14(1):68–80.
21. Swayne DF, Lang DT, Buja A, Cook D. GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. Comput Stat Data Anal. 2003;43(4):423–44.
22. Dimitriadou E, Hornik K, Leisch F. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-27. 2011;
23. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.
24. Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics. 2013;29(16):1925–33.
25. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: Challenges and opportunities. Brief Bioinform. 2016;17(1):23–32.
26. Bååth R. The State of Naming Conventions in R. R J. 2012;4(2):74–5.
27. Peng RD. Reproducible Research in Computing Science. Science (80-). 2011;334(6060):1226–7.