Crowdsourcing and Its Applications on Data Mining : A Brief Survey

K. Karthika, R. Durga Devi
Department of Computer Applications, Saradha Gangadharan College, Puduchery, Tamil Nadu, India

ABSTRACT

Crowdsourcing allows large-scale and flexible invocation of human input for data gathering and analysis, which introduces a new paradigm of data mining process. Traditional data mining methods often require the experts in analytic domains to annotate the data. However, it is expensive and usually takes a long time. Crowdsourcing enables the use of heterogeneous background knowledge from volunteers and distributes the annotation process to small portions of efforts from different contributions. This paper reviews the state-of-the-arts on the crowdsourcing for data mining in recent years. We first review the challenges and opportunities of data mining tasks using crowdsourcing, and summarize the framework of them. Then we highlight several exemplars works in each component of the framework, including question designing, data mining and quality control. Finally, we conclude the limitation of crowdsourcing for data mining and suggest related areas for future research.

Keywords: Clustering, Crowdsourcing, Data mining, Sampling, Quality control

I. INTRODUCTION

People from different fields analyze a variety of datasets to understand human behaviors, find new trends in society, and possibly formulate adequate policies in response. Typically, we address the problem of finding interesting and unknown patterns via data mining methodology. Data mining enables people to extract information from a data set and convert it into a comprehensible structure for further use. Typical data mining techniques, however, are not suitable for current applications. First, when mining the datasets, we must have access to all relevant information. In fact, it is impossible to obtain all these transactions, which mainly because of the properties of the human memory.

People's memories are prone to remember summaries, rather than exact details [1]. Consider the following case. A social scientist wants to analyze life habits of people. The database includes leisure activities (watching TV, jogging, reading, etc.) correlated with time of the day, weather and so on. But it is unrealistic for people to recall an exhaustive list of all cases they did. People can make assumptions in order to compensate the loss of information by crowdsourcing the mining task. Second, some mining algorithms are time-consuming, especially used for large datasets, which also leads to much more extra cost. Finally, raw data mining technologies are lack of related information. Algorithm has to be taught the knowledge before mining. For example, for the classification problem, labeled data is used for training the classifier to have the ability of classifying new coming test data. However, acquiring the labeled data is time consuming and costly. In the circumstances, we can solve this problem by crowdsourcing.

As crowdsourcing is based on the people who have the incentives to work on small tasks, the mining tasks can benefit from the aggregation of labeling work which is time-controllable, flexible, easy to implement due to the current crowdsourcing platform. Crowdsourcing is an emerging and powerful information procurement paradigm that has appeared under many names, including social computing, collective intelligence and human computation [2]. Requesters decompose the whole task into several small tasks and push them to the crowd, and workers accomplish questions for intrinsic or extrinsic reasons. Although people may not remember all of transactions precisely, many current studies prove that simple summaries can still achieve a positive result, and even more complicated questions. Crowdsourcing has played important roles in data mining. In some kind of scenarios, it can help people resolve the problems in a more efficient way and give them deeply understanding to apply crowdsourcing. Here we give some situations for the applications of crowdsourcing techniques in various real-world data mining tasks. Crisis Map: Crisis map is one of the most representative applications of crowdsourcing. It is a platform, designed to do information collection, analysis of mass data and display in a straightforward way in real time during a crisis. It has become a powerful mechanism for a large number of people to contribute about crisis events.

II. BACKGROUND DETAILS

Framework:
Traditional data mining methodologies and technologies are sometimes time-consuming, inflexible, expensive to implement, and poor scalable. Crowdsourcing can be applied to manage data and extract interesting patterns from the data sets more efficiently and intelligently by comparison. From existing work of crowdsourcing techniques we conclude that using crowdsourcing for data mining can be performed by following a three-step procedure: question design, mining and quality control.

Question Design:
Well-designed tasks can obtain high-quality answers. Questions should be designed based on the purpose of the data mining task. We address the problem of effective crowdsourcing, namely gathering data from the crowd in a way that is economical in time and expense.

Mining:
The mining phase absolutely takes the center stage in the whole process. Data mining tasks can be divided into the multiple kinds: classification, clustering, semi-supervised learning, and association rules mining. Classification has been widely used in many fields, such as face recognition, disaster rescue. Some research takes advantages of crowdsourcing to identify association rules between relating signs and symptoms to diseases. Crowdsourcing appears to have several important merits compared with other automatic knowledgebased approaches.

Quality Control:
Due to the nature of crowdsourcing task, a quality control step is necessary for the result after the mining step. Malicious workers, who are only attempting to maximize their income or lack of necessary training, are detrimental to the mining result. Quality control step uses vote system, redundant workers, worker's reputation and other methods to pick out the irresponsible workers. In the following sections, we will discuss these three steps, respectively.

III. MINING DATA FROM CROWDSOURCING

Various types of data mining tasks can be accomplished by means of crowdsourcing, e.g. classification, clustering, semi supervised learning, and association rule mining. Traditional algorithms have difficulties in tackling these problems, for the lack of knowledge. In these situations, the powerful crowds can perform more accurately, flexibly and efficiently than the existing automatic algorithms. We will discuss how crowdsourcing can be used to solve these problems and provide some application scenario.

3.1 Classification

Crowdsourcing can be utilized to solve classification problem. Crowdsourcing method has much more advantages over the general data mining technology. A typical classification problem is to distinguish male and female from a social network user. In the original data mining perspective, a classifier is built to extract features from the given datasets in the first step. In the second step, the classifier is used for classification. Compared to this, if we distribute this task to the crowd, everyone can classify the objects immediately. Sometimes we can say that the wisdom of the crowd allows for more accuracy than any other classification algorithm. Documents categorization can be accomplished by users on the website as well. This approach has been applied on many domains successfully, such as Digg and Yahoo! Directory. Digg is an aggregator to customize the user's news front page. Digg also allows users to tag to the submitted links, which widen the scope to include more relevant articles the users may be interested in. We find that accuracy increases as the more and more users participate in. Another widely known example is CAPTCHA.

''CAPTCHA is a challenge response test used on the World Wide Web to determine whether a user is a human or a computer''. Technologies at present cannot recognize distorted text as fast as humans can. Human enter the characters to digitize handwriting text . present an algorithm, named CASCADE, to create an overall consistent taxonomy by distributing HITs to many individuals, each of whom has only a partial view of the data. CASCADE hires many unskilled labors to produce taxonomies. The quality of classification is approximate to that of human experts, while the cost of CASCADE is very cheap. Furthermore, present DELUGE, an improved workflow on the basis of CASCADE. DELUGE produces taxonomies with equivalent quality in spite of reducing the workforce. The categorization step, which is the most consuming, is optimized by decision theory. The experiment result demonstrates that less than 10% of the workers are required by the original approach. As we mentioned before, crowdsourcing helps a lot in disaster rescue, establish an online framework that generates human computation resources to tackle an image labeling task, classifying post-disaster photos according to damage extent. In real life, such type of information is needed to manage risks in disaster-prone areas, both in pre-disaster risk reductions and post-disaster damage assessments. Other related works that considering the budget allocation contributed to the classification work as well raise the issue of how to allocate the budget for redundant workers when dealing with classification tasks, where the key challenge is to find a proper balance between the total cost and quality. They propose CrowdBudget, a budget allocation algorithm, aiming to minimize estimation error with the limited fund.

3.2 Clustering

Clustering is more complicated than classification problem. One of the aspects, there are lots of ways to define the similarity between items. Different measure of similarity may lead to different result. Similarly, we can perform cluster task on the crowdsourcing platform. Many recent social networking sites give humans permission to create categories. In the case of Twitter, users assign tags to their tweets in order to follow up the trending topics. This facilitates quick retrieval when searching for tweets and again, and forms a discussion groups about the hot issue automatically. What is more, a large set of Tweets relevant to a particular cluster can be an excellent source for professionals to analyze. Create a friendly environment, using crowd to visualize web images into clusters. The method has two stages. The first stage separates an image set into multiple clusters and the second stage purifies each generated cluster independently. During the whole stage, computers select informative images and the crowds help to label the images to improve the quality. The experimental results demonstrate the combinations of computers and a large number of human workers benefit high-quality visual clusters. Here we require addressing some challenges in crowdsourced clustering.

1. Each worker has only a fraction of the data, so we need additional algorithm to merge the results.
2. Different workers may have different clustering standard, leading to produce different numbers of categories.
3. The underlying category structure may be hierarchical. According to the intractable problems mentioned above, propose a model, based on variationally bayes method, of how crowdsourcing can be applied to clustering. First, divide the dataset into overlap subsets. And then workers propose partial clustering. Finally, use Bayes model to aggregate the partial clusters into one cluster.

3.3 Semi-supervised learning

In semi-supervised learning, we use labeled data to acquire necessary knowledge, and then label the unlabeled data. This procedure significantly increases the learning accuracy. Similar to previous tasks, semi-supervised learning can also be performed with crowdsourcing. aim to achieve more accuracy when inferring consensus labels, with correspondingly less labeled training data for estimating worker accuracy by a Naive Bayes approach. We can apply this method in the situation when we have large amount of unlabeled items and a very small set of expert labeled items. As human have better learning ability than the algorithm, requesters can provide essential knowledge of how the given task can be performed correctly. This is a well-practiced labeling technique for sophisticated labeling tasks in the data mining field

3.4 Sampling

Sampling is in correlation with the selection of a subset with sufficient information so that people can easily verify hypotheses devised from the sample information in the whole datasets. And it is one of the complex tasks in data mining. A challenging problem to be solved here is how the requester should select the appropriate distribution so that benefit of the information gathered from the sample is maximized. Crowds have been proved to be trustworthy data samplers, and they keep working on enhancing the precision of the results in maximally informative samples

3.5 Association rule mining

Data mining techniques have been developed for discovering and identifying underlying association rules among data items. The typical application is shopping baskets analysis. That is, a market analyst can explore relation about which items are purchased together by analyzing purchasing records. However, when referring to human behavior, it is impossible to get access to all the transactions. This is because, typically, the everyday actions of people are not recorded in detail, except in their own memories, which are limited in terms of exact recollection. Indeed, social studies show that instead of full details, people often tend to recall sufficient information in the form of summaries when asked the appropriate questions. lay the foundations of crowd mining for the first time. They define the basic concepts of mining the association rules. Then, they present an integrated system consisted of general-purpose components, incorporating interactive selection of questions to ask, effective mining component, error estimation and so on. Another article from present a demo named CrowdMiner. The essence of Crowd Miner is an algorithm enabling the mining of appealing data patterns from the crowd. It allows flexible choice about appropriate questions to ask the crowd as well, with the purpose of gathering more information with fewer questions.

3.6 Validation

Similarly, we can perform task to human to validate the correctness of mining algorithm and predict the mining result of the automated method on large dataset [3]. [4] want to identify influential bloggers at a blog site. As we know, there is no training and testing data for them to evaluate the efficiency of the proposed model. They use crowdsourced result generated on Digg as a reasonable reference to compare with their automatic techniques. The crowdsourcing results validate their hypothesis.

IV. QUALITY CONTROL

Crowdsourcing is a powerful platform for many mining tasks. Workers may misunderstand the tasks, make mistakes, or deliberately cheat the system, which can cause errors or bad results. Bad answers consume a lot of time and money to filter them out. The reason is twofold. On one hand, many workers fulfilled tasks to kill time or gain sense of achievement in the beginning, with the payment being only a minor attraction. Nowadays, the overwhelming majority of workers are attracted by the financial reward [5]. Payment is the easiest method to motivate people. However, monetary incentives can effectively increase participation, but cannot improve quality. As a consequence, a high number of malicious users arise. They try to finish HITs as quickly as possible in order to maximize their profit. This leads to a mass of generic answers. One the other hand, the worker may lack expertise or skills to handle some kind of complex job [6]. Incorrect answers may be provided because of this. To tackle this problem, we can provide some basic knowledge to workers before the work, or require some qualifications to prove themselves qualified to finish the task. In a nutshell, the quality of crowdsourced data has great influence on the mining result, so researchers have to pay careful attention to it. To improve the trustworthiness of mining result, various techniques could be employed. They are discussed as follows, respectively.

4.1 Vote

Voting system hires additional spammers to judge the crowdsourcing outcome, and follows majority rule, which is simple to implement in real world application. Voting system is quite successful and easy-implement in determining the credibility of messages on the web. In social media sites, people use thumb up or thump down to express their attitudes for or against. For instance, on YouTube, users can provide will hide the comments with too many negative feedbacks automatically. Another example is eBay, the buyer votes to seller to give other buyers a reference to the product. And the seller vote according to the buyer’s trustworthiness. Although voting approach has its advantages, we have to admit this approach does have drawbacks. Minority voters have less access to express their views and so researchers would be less likely to benefit from these special ideas.

4.2 Redundant Work

Apparently, redundant work means the requester hire redundant workers to finish the same crowdsourcing task. Actually, redundancy work is widely used to identify the correct answers by the requesters. However, what we should pay attention to is that redundancy is not a panacea. Large-scale redundancy is expensive, and redundant workers sometimes may not lead to good result. Therefore, we can apply redundant workers to controversial item to save money, not to all the items.

4.3 Worker's Reputation

Worker’s reputation is composed primarily of the worker’s accuracy on previously submitted HITs. Reputation is a practical judgment on workers’ trustworthiness, which urges people to complete work with high quality continuously. It has become a popular method for evaluation of the quality of workers not only on crowdsourcing platforms, but also on a lot of online forums In Amazon Mechanical Turk, requesters can require the level of worker’s HITs Approval Rate and some other qualifications, such as language skill. When cheating is detected, the reputation reduces and the system forbids low reputation workers, e.g. users who fail two tasks may be put into blacklist [7]. [8] propose a reputation management framework, which adequately takes into account the values of the tasks completed, the trustworthiness of the assessors, the results of the tasks and the time of evaluation in order to achieve more credible quality metrics for workers and assessors. [9] propose an algorithm that improve the existing advanced techniques of the labeling process in crowdsourcing platform and can be applied when the workers should answer a multiple-choice question to complete a task. The algorithm enables the separation of intrinsic error rate from the bias worker. Finally, the algorithm produces a scalar score to measure the intrinsic quality of each worker. Worker’s reputation has potential problems. First, the major determinant factor of human’s reputation is the acceptance rate of HITs. The requesters always accept all answers and do not dispose the noisy data right now. Afterwards, requesters do not give feedback to the workers, respectively. The malicious users take advantages of the loophole to receive the increase in reputation and start to complete next tasks. Second, reputation system could not avoid cheating. It creates a scam, which is designed to accelerate the reputation improvement, named rank boosting on his weblog. With this strategy the worker creates a requester account, distributes a number of simple HITs and immediately completes them with his worker account. The worker almost spends no money in boosting his rank. 5.4. Gold standard Gold standard is the benchmark that is the best available in particular situation. It does not have to be necessarily the best answer for the condition in giving terms. In crowdsourcing domain, we may pay expert to use small group of data to set gold standard, and then we can compare the gold standard data with the HIT‘s work to judge the reliability and filter out the poor-quality workers. [10] build a system on the idea of gold standard questions that the requester has labeled as the ground truth. When encountering the ground truth questions, the worker’s answer must include at least one choice from an inclusion list and none from an exclusion list. [11] insert gold standard data into questions and robustly rejected bad answers to ensure quality when workers made mistakes in those gold standard data. [12] extend gold standard questions with novel technique. Also, some trap questions can be mixed with real questions and the system can easily notify the bad answers. Moreover, [13] suggest that the tasks should be well-priced and make clear enough for workers to follow. They also propose several approaches to restrain cheating, such like using images of sentences instead of text in order to prohibit copying and pasting in translation tasks. For some situations, like kinds of creative jobs, gold standard seems difficult to set.

V. CONCLUSION

This paper reviews recent work on crowdsourcing-based data mining techniques. Crowdsourcing can do data mining and extract addition information from the datasets more efficiently and intelligently than traditional methods. It has to deal with lots of challenges like the low quality of answers from the crowds to apply crowdsourcing to data mining. In this paper, we point out these challenges and introduce the general procedures of an integrated data mining task in crowdsourcing. The task is often partitioned into three phases: question designing, data mining and quality controlling. We take a deep overview of work in each phase and conclude their contributions. Besides those discussed in Section 7, there are some promising directions for future research, such as designing crowdsourcing platform especially for governments or companies, algorithms in various steps to process large scale mining task, the background system for requesters to perform detailed analysis. Another future research task is to apply crowdsourcing to discovery knowledge for.

VI. REFERENCES

[1]. Boim, R., Greenshpan, O., Milo, T., Novgorodov, S., Polyzotis, N., & Tan, W. C, Asking the right questions in crowd data sourcing, Proc. 28th international conference on data engineering (ICDE), 2012.

[2]. Quinn, A. J., & Bederson, B. B, Human computation: A survey and taxonomy of a growing field, Proc. SIGCHI conference on human factors in computing systems, 2011.

[3]. Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H, Maximizing benefits from crowdsourced data, Computational and Mathematical Organization Theory, 18(3), 2012, 257–279.

[4]. Agarwal, N., Liu, H., Tang, L., & Yu, P. S, Identifying the influential bloggers in a community, Proc. international conference on web search and data mining, 2008.

[5]. Eickhoff, C., & de Vries A, How crowdsourceable is your task, Proc. fourth ACM international conference on web search and data mining (WSDM), 2011.

[6]. Liu, X., Lu, M., Ooi, B. C., Shen, Y., Wu, S., & Zhang, M, Cdas: A crowdsourcing data analytics system, Proc. VLDB Endowment, 2012.

[7]. Heimerl, K., Gawalt, B., Chen, K., Parikh, T., & Hartmann, Community Sourcing: Engaging local crowds to perform expert work via physical kiosks, Proc. ACM annual conference on human factors in computing systems,2012.

[8]. Allahbakhsh, M., Ignjatovic, A., Benatallah, B., Beheshti, S. M. R., Bertino, E., & Foo, Reputation management in crowdsourcing systems, Proc. 8th international conference on collaborative computing: networking, applications and worksharing,2012.

[9]. Ipeirotis, P. G., Provost, F., & Wang, J, Quality management on Amazon mechanical Turk, Proc. ACM SIGKDD workshop on human computation, 2010.

[10]. Bernstein, M. S., Teevan, J., Dumais, S., Liebling, D., & Horvitz, Direct answers for search queries in the long tail, Proc. SIGCHI conference on human factors in computing systems, 2012

[11]. Le, J., Edmonds, A., Hester, V., & Biewald, L., Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution, Proc. SIGIR 2010 workshop on crowdsourcing for search evaluation,2010.

[12]. Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces, Proc. 24th annual ACM symposium on user interface software and technology, 2011.

[13]. Callison-Burch, C., & Dredze, M, Creating speech and language data with Amazon’s mechanical Turk, Proc. NAACL HLT 2010 workshop on creating speech and language data with Amazon’s mechanical Turk, 2010.