Entropy **2014**, *16*, 3866-3877; doi:10.3390/e16073866

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Many Can Work Better than the Best: Diagnosing with Medical Images via Crowdsourcing

Xian-Hong Xiang [1], Xiao-Yu Huang [2,3], Xiao-Ling Zhang [1], Chun-Fang Cai [4], Jian-Yong Yang [1,*] and Lei Li [3]

[1] Department of Interventional Radiology, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510080, China; E-Mails: med.interventional@163.com (X-H.X.); zhangxiaoling_77@163.com (X.-L.Z.)

[2] School of Economics and Commerce, South China University of Technology, Guangzhou 510006, China; E-Mail: echxy@scut.edu.cn

[3] Software Institute, Sun Yat-Sen University, Guangzhou 510275, China; E-Mail: lncsri07@mail.sysu.edu.cn

[4] Department of Gynaecology and Obstetrics, Guangzhou Women and Children Medical Center, Guangzhou 510623, China; E-Mail: dmdata@126.com

* Author to whom correspondence should be addressed; E-Mail: cjr.yangjianyong@vip.163.com; Tel.: +86-20-87755766.

Received: 8 March 2014; in revise form: 22 June 2014 / Accepted: 3 July 2014 / Published: 14 July 2014

Abstract: We study a crowdsourcing-based diagnosis algorithm, which is against the fact that currently *we do not lack medical staff, but high level experts*. Our approach is to make use of the general practitioners' efforts: For every patient whose illness cannot be judged definitely, we arrange for them to be diagnosed multiple times by different doctors, and we collect the all diagnosis results to derive the final judgement. Our inference model is based on the statistical consistency of the diagnosis data. To evaluate the proposed model, we conduct experiments on both the synthetic and real data; the results show that it outperforms the benchmarks.

Keywords: medical images based diagnosis; crowdsourcing; entropy; Kullback–Leibler divergence

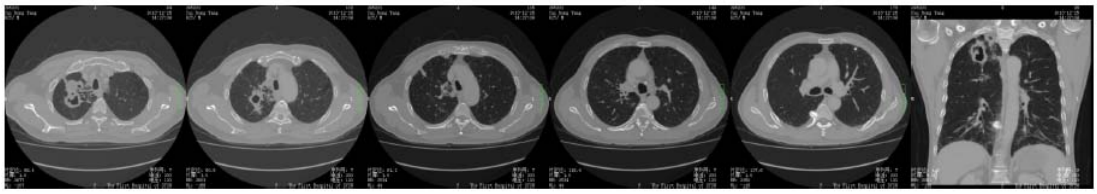
---

1. Introduction

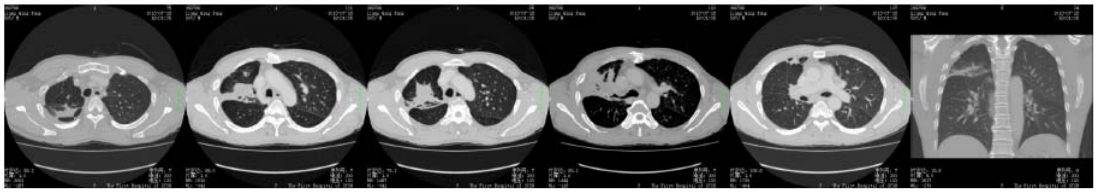
Image based diagnosis has been developed and widely used in medical fields for decades, according to some statistics, up until 2010, 5 billion medical imaging studies had been conducted worldwide [1]. By analyzing a great deal of information yielded through imaging techniques such as X-ray Computerized Tomography (CT) and Magnetic Resonance Imaging (MRI), doctors reveal, diagnose, or examine disease for patients, examples including strokes [2] and cancers [3].

However, as well known, accurate analysis and interpretation of a medical image relies heavily on the knowledge and experience of experts. In cases of images that contain many details, the diagnosis process might become tedious and time consuming even for well-trained professionals. For illustration, let us study some CT images presented in Figure 1, where the first row is from some lung cancer patients, and the second row is of the pulmonary abscess patients. We see the differences between the two rows of images are very subtle, so it is not easy to distinguish the two kinds of the patients from each other, even for some well educated junior doctors. For example, as mentioned below, we have recruited 13 graduate students to diagnose 50 patients according to their CT images, we find the result is far from optimistic: On average, every student only has 19 correct diagnoses.

Figure 1. Some sample medical images, pictures in the first row are from the pulmonary abscess patients, in the second row are from the lung cancer patients. (**a**) Some sample images of the pulmonary abscess patients; (**b**) Some sample images of the lung cancer patients.



(**a**)



(**b**)

Meanwhile, a brutal reality is true worldwide: well trained experts are rare. For example, according to some public reports, in China, until July of 2012, among the 1.3 billion population there were about 2 million doctors. In other words, the number of doctors per 1000 people was around 1.5. Although this is greater than 1.25, which is suggested as the lower bound by WHO (World Health Organization), only half of the doctor population holds a bachelor or higher degree in medicine. What is more, among the the total doctor population, the ratio of people with a master or higher degree is below 8%. In addition, because of various objective conditions such income and life, almost all the well-trained and experienced

doctors gather in a very few highly developed cities of China, such as Beijing [4], Shanghai [5] and Guangzhou [6]. Almost all of the other cities suffer from a severe shortage of high level doctors.

Because of the unbalanced distribution of the experts, for many hospitals, when they have patients who cannot be definitively diagnosed, they often need to ask for help from outside experts. This approach, despite the inefficiency and the extra cost, does not always work well, because the experts are often needed by their own business.

In the present article, we propose another attempt to approach the shortage of experts with respect to the context of medical image based diagnosis. Our basic idea is regarding the observation that what we lack are the *experts*, but not the *general practitioners*, so it is possible to release the experts from endless requests via making use of the general practitioners's efforts. Our solution is the *crowdsourcing* [7] scheme, the details of which are presented in Table 1.

Table 1. The working scheme of the crowdsourcing based diagnosis.

1.	For every patient who can not be diagnosed definitely, do:
2.	Invite some other doctors to diagnose the patient based on her medical images;
3.	Summarize the all diagnosis results and make the final decision;
4.	End

At first glance, the procedure of Table 1 looks very much like the expert consultation (ES) system. However, there are some fundamental differences between the two approaches: Firstly, the ES scheme often requires the participation of experts, while the crowdsourcing scheme only needs the general practitioners (but of course, experts are welcome.). Secondly, in the ES scheme, all the experts usually take part in the diagnosis together and achieve the unique conclusion in the end. In the crowdsourcing scheme, every doctor works independently and the final judgement is derived by some algorithm that takes the all doctors' conclusions as input.

Our contribution is three fold, the summarization is as follows:

(1) We propose the crowdsourcing based diagnosis paradigm;
(2) We present a statistical consistency based learning algorithm, which ensembles all the doctors' diagnosis conclusions and derived the final decision;
(3) We evaluate the proposed approach with the synthetic and real data.

The remainder of the paper is organized as follows. Section 2 discusses the related works on crowdsourcing; Section 3 describes a real medical image based diagnosis results set that is used in the work; in Section 4, we present our crowdsourcing based diagnosis method; Section 5 is devoted to the experiments, and Section 6 is the conclusion.

2. Crowdsourcing

To the best of our knowledge, the term *crowdsourcing* was first proposed by Jeff Howe as the composition of the terms "wisdom of crowds" and "outsourcing" [8,9]. In essence, crowdsourcing is one type of Human as a Service (HuaaS), where a group of (not necessary expert) people (or

workers) are asked to do a task of that often needs professional background, such as natural language processing [10], movie recommendation [11], optical character recognition [12], image classification [13,14] and dermatology research [15]. One of the most famous crowdsourcing examples is *Wikipedia*, where thousands of users contributes the creation of the world's largest encyclopedia every day. And some other well known instances include the Amazon Mechanical Turk platform [16], the the Galaxy Zoo project [17] and the Click Worker project [18].

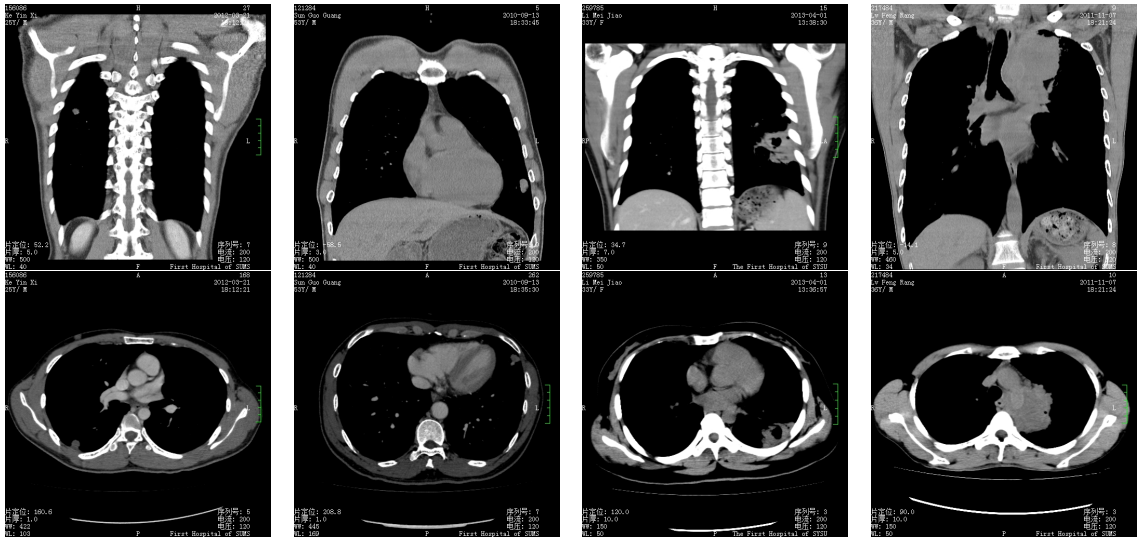
Since most of the crowdsourcing contributors are not domain experts, so their working results are often of relatively low quality. Hence, naturally, a central concern of crowdsourcing is *How to combine the individuals' works to derive high quality results.* The approaches, roughly speaking, can be categorized into two classes: The first category is the data content independent (DCI) method, where the ensemble algorithm only takes the individuals' conclusions as input and makes the final judgement. Among all the DCI methods, the most used one is the majority voting algorithm [10], which suggests that for every item of the task, the ground truth is the one that is elected by the most workers. Despite its simplicity, the majority voting algorithm is well recognized as the most stable one among various crowdsourcing algorithms [7,19,20], and achieves surprising success in many crowdsourcing applications. However, for the naive majority voting algorithm and its variations, almost all of them need every task to be done multiple times by different workers. This requirement, when the actual cost is taken into account, is infeasible in many real applications. Addressed to the shortage of the DCI methods, as the second category algorithm, the data content dependent (DCD) method is proposed. A typical DCD policy usually consists of two stages: In the first stage, it learns the behavior of the workers from their working results, *i.e.*, for every worker w_i, it treats the items worked by w_i along with the working results as the training data and learns the predictor to simulate the behavior of w_i, then applies the learned model to act as w_i to make predictions on the other items [21,22]. In the second stage, the algorithm ensembles the all working results (both of the workers' results and the prediction results) and makes the final judgements. To avoid the undertraining problem, the DCD approach often requires every worker to have large enough working results.

3. Data

The dataset we use is composed of 50 patients' CT medical images, for every patient there are 300–400 images. Every patient is in one of the four categories: pulmonary tuberculosis, lung cancer, pulmonary abscess and pulmonary metastasis, some samples of the images are presented in Figure 2.

We recruit 13 volunteers to diagnose (or to *label*) the patients according to their images, all the volunteers are 2nd or 3rd year graduate students of the medical imaging major. We ask every student to make their diagnosis for every patient according to the images independently. The average accuracy of the volunteers is 39.54%, *i.e.*, on average, every student only has 19 accurate diagnosis. Besides, the best volunteer achieved an accuracy of 50%, while the worst one only has a accuracy of 20%.

Figure 2. Some sample medical images, (**a**) pulmonary tuberculosis; (**b**) lung cancer; (**c**) pulmonary abscess; (**d**) pulmonary metastasis.



(**a**) (**b**) (**c**) (**d**)

4. Method

In our problem, every worker only labels 50 patients, while for every patient there are more than 300 high resolution medical images, so it is easy to become trapped into the undertraining dilemma if we try to learn the worker's behavior via their working results. As a result, we take the DCI policy to make diagnosis judgement.

Our idea is based on the statistical consistency of the patients' diagnosis results: Denote the set of available doctors as $\{D_1, D_2, \ldots, D_n\}$, the set of patients as $\{P_1, P_2, \ldots, P_m\}$, the set of possible illnesses (or, labels) as $\{I_1, I_2, \ldots, I_k\}$. Let S_i represent the set of diagnosis results of P_i (Throughout this paper, unlike conventional definitions, we allow a set to contain duplicate values.). We use $\mathfrak{D}_i$ to represent the distribution on S_i. Specifically, we denote $S_0 = S_1 \cup S_2 \cup \ldots S_n$, and $\mathfrak{D}_0$ the distribution on S_0. It is noteworthy that here we do not require the patients to be diagnosed by the all doctors, hence, the distributions $\mathfrak{D}_i$s are estimated only by the collected diagnosis data. For patient P_i, to determine which illness she has, our idea is to choose the one from $\mathcal{I}$ which leads to the minimal changes to both the global distribution $\mathfrak{D}_0$ and individual distribution $\mathfrak{D}_i$.

4.1. Preliminaries

Throughout this paper we use upper case letters (e.g., $X, Y, Z, \ldots$) to denote the random variables, and lower cases to represent the instances.

Our work is mainly based on Information Theory. Below we introduce some definitions and preliminary results used in this paper. Most of them can be found in [23].

Let $\mathcal{P}$ be a distribution with $p(X)$ as the probability density function (p.d.f) for $X \sim \mathcal{P}$, then *entropy* of X is defined as

$$H(X) = -\int p(x) \ln p(x) dx.$$

Given a distribution $\mathcal{Q}$ with $q(X)$ as the p.d.f, we employ the Kullback–Leibler divergence to measure the distance between $\mathcal{P}$ and $\mathcal{Q}$, which is defined as

$$KL(P||Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx.$$

Our assumption of the proposed algorithm is as follows:

Assumption 1. *For* $i \in \{0, 1, 2, \ldots, n\}$, *the distribution* $\mathfrak{D}_i$ *is multinominal with probability* $\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}$.

For every $0 \leq i \leq n$, we use $n_{i,j}$ to denote the number of I_j in S_i, let $n_{i,0} = \sum_{l=1}^{k} n_{i,l}$, we have the following theorem:

Theorem 1. *Let* $p^*_{i,1}, p^*_{i,2}, \ldots, p^*_{i,k}$ *be the solution to the following problem*

$$\{p^*_{i,1}, p^*_{i,2}, \ldots, p^*_{i,k}\} = \underset{\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}}{argmax} \quad Pr(S_i) \tag{1}$$

then for $j = 1, 2, \ldots, k$

$$p_{i,j} = \frac{n_{i,j}}{n_{i,0}}. \tag{2}$$

Proof of Theorem 1 According to Assumption (1),

$$Pr(S_i) = \frac{n_{i,0}!}{n_{i,1}! n_{i,2}! \ldots n_{i,k}!} \prod_{l=1}^{k} p_{i,l}^{n_{i,l}} \tag{3}$$

Take logarithm on both sides of the equation above, we have

$$\ln Pr(S_i) = \ln \frac{n_{i,0}!}{n_{i,1}! n_{i,2}! \ldots n_{i,k}!} + \sum_{l=1}^{k} n_{i,l} \ln p_{i,l}. \tag{4}$$

Noting that the term $\ln \frac{n_{i,0}!}{n_{i,1}! n_{i,2}! \ldots n_{i,k}!}$ is a constant and

$$\sum_{l=1}^{k} p_{i,l} = 1. \tag{5}$$

Let

$$T_i(p_{i,1}, p_{i,2}, \ldots, p_{i,k}) = \sum_{l=1}^{k} n_{i,l} \ln p_{i,l} + \lambda(1 - \sum_{l=1}^{k} p_{i,l}).$$

where $\lambda > 0$ is fixed, then problem of Equation (1) is equivalent to the following:

$$\{p^*_{i,1}, p^*_{i,2}, \ldots, p^*_{i,k}\} = \underset{\{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}}{argmax} \quad T_i(p_{i,1}, p_{i,2}, \ldots, p_{i,k}). \tag{6}$$

For $l = 1, 2, \ldots, k$, let

$$\frac{\partial T_i(p_{i,1}, p_{i,2}, \ldots, p_{i,k})}{\partial p_{i,l}} = 0,$$

we have

$$p_{i,l} = \frac{n_{i,l}}{\lambda}. \tag{7}$$

plug Equation (7) into Equation (5), we achieve $\lambda = n_{i,0}$, hence we have the proof. □

4.2. Diagnosis with Crowdsourcing

Given the sets of the diagnosis results $S_1, S_2, \ldots, S_n$, we pretend there is an extra *oracle doctor* to make the final judgement for every patient. Denote the illness of P_i given by the oracle as O_i. For an arbitrary $X \in \{I_1, I_2, \ldots, I_k\}$, noting that the distribution on $S_i \cup \{X\}$ will always differ from that on S_i, denote the distribution on $S_i \cup \{X\}$ as $\mathfrak{D}_i^{new}$, we assume O_i is the one that most consistent with their existing diagnosis, or, formally,

$$O_i = \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmin} \quad KL(\mathfrak{D}_i || \mathfrak{D}_i^{new}) \tag{8}$$

Denote $Pr_i(\cdot)$ as the probability function of $\mathfrak{D}_i$ and $Pr_i^{new}(\cdot)$ the function of $\mathfrak{D}_i^{new}$, noting that:

$$\frac{1}{|S_i|} \ln \frac{Pr(S_i)}{Pr^{new}(S_i)} = \frac{1}{|S_i|} \ln \frac{\prod_{Z \in S_i} Pr(Z)}{\prod_{Z \in S_i} Pr^{new}(Z)} \tag{9}$$

$$= \frac{1}{|S_i|} \sum_{Z \in S_i} \ln \frac{Pr_i(Z)}{Pr_i^{new}(Z)} \tag{10}$$

$$= \frac{1}{|S_i|} \sum_{I_j \in S_i} \sum_{Z = I_j} \ln \frac{Pr_i(Z)}{Pr_i^{new}(Z)} \tag{11}$$

$$= \sum_{I_j \in S_i} \frac{n_{i,j}}{n_{i,0}} \ln \frac{Pr_i(I_j)}{Pr_i^{new}(I_j)} \tag{12}$$

According to Theorem 1, $Pr_i(I_j) = \frac{n_{i,j}}{n_{i,0}}$, hence, the last equation above is exactly the divergence $KL(\mathfrak{D}_i || \mathfrak{D}_i^{new})$.

Now we seek the solution to Equation (8), firstly, for $X = I_l (1 \leq l \leq k)$, we have

$$p_{i,j}^{new} = \begin{cases} \frac{n_{i,j}}{n_{i,0}+1} & \text{if } j \neq l, \\ \frac{n_{i,j}+1}{n_{i,0}+1} & \text{if } j = l. \end{cases}$$

Noting that in Equation (10), the term $\frac{1}{|S_i|}$ is fixed, so

$$KL(\mathfrak{D}_i || \mathfrak{D}_i^{new}) \propto \ln \frac{Pr_i(S_i)}{Pr_i^{new}(S_i)} \tag{13}$$

$$= \ln Pr_i(S_i) - \ln Pr_i^{new}(S_i) \tag{14}$$

Since $\ln Pr_i(S_i)$ is a constant, the target Equation (8) is equivalent to the following:

$$O_i = \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmax} \quad \ln Pr_i^{new}(S_i) \tag{15}$$

where for $X = I_l$,

$$\ln Pr_i^{new}(S_i) = \sum_{j=1}^{k} (n_{i,j} + \mathbb{I}(j = l)) \ln p_{i,j}^{new} \tag{16}$$

In addition to Equation (8), it's noteworthy that the introduction of O_i will also lead to changes to the global distribution $\mathfrak{D}_0$, these changes, should be as small as possible, too. Therefore, similar to Equation (8), we have:

$$O_i = \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmin} \quad KL(\mathfrak{D}_0 || \mathfrak{D}_0^{new}) \tag{17}$$

where we use $\mathfrak{D}_0^{new}$ to denote the distribution on $\mathfrak{D}_0 \cup \{O_i\}$.

Denote $Pr_0(\cdot)$ as the probability function of $\mathfrak{D}_0$ and $Pr_0^{new}(\cdot)$ the function of $\mathfrak{D}_0^{new}$, analog to the procedure above, we also have:

$$O_i = \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmax} \quad \ln Pr_0^{new}(S_0) \tag{18}$$

where for $X = I_l$,

$$\ln Pr_0^{new}(S_0) = \sum_{j=1}^{k} (n_{0,j} + \mathbb{I}(j = l)) \ln p_{0,j}^{new} \tag{19}$$

With Equations (16) and (19), we have the final decision target:

$$O_i = \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmax} \quad \ln Pr_i^{new}(S_i) + \lambda \ln Pr_0^{new}(S_0) \tag{20}$$

$$= \underset{X \in \{I_1, I_2, \ldots, I_k\}}{argmax} \quad \sum_{j=1}^{k} (n_{i,j} + \mathbb{I}(j = l)) \ln p_{i,j}^{new} + \lambda \sum_{j=1}^{k} (n_{0,j} + \mathbb{I}(j = l)) \ln p_{0,j}^{new} \tag{21}$$

where $\lambda > 0$ is the tradeoff factor,

4.3. Algorithm

To ensemble the individuals' judgements and make the final diagnosis for the patients, we adopt the enumeration policy, *i.e.*, for every patient, we enumerate the all possible illnesses and calculate the target values respectively. We take the one with the minimum value of Equation (20) as the final judgement of the patient.

The details of the algorithm are presented in Algorithm 1, where line 6 is from Equation (21). We see in the algorithm there nk iterations, besides, in each iteration, to calculate the probability $p_{i,j}s$ ($0 \leq i \leq n$, $1 \leq j \leq k$), we need at most m scans to count the diagnosis given by the m doctors to the patient, so the time complexity of the algorithm is $\theta(nmk)$.

Algorithm 1: Diagnose via Crowdsourcing.

Input: Candidate illness set $\{I_1, I_2, \ldots, I_k\}$, doctor set $\{D_1, D_2, \ldots, D_m\}$, patient set $\{P_1, P_2, \ldots, P_n\}$ and the diagnosis sets $S_1, S_2, \ldots, S_n$, initial value of λ.

Output: The judgements $O_1, O_2, \ldots, O_n$, where O_i corresponds to P_i.

```
1  for i=1 to n do
2      max_val = −∞;
3      O_i = null;
4      for j=1 to k do
5          X = I_j;
6          temp = Σ_{j=1}^{k} (n_{i,j} + 𝕀(j = l)) ln p_{i,j}^{new} + λ Σ_{j=1}^{k} (n_{0,j} + 𝕀(j = l)) ln p_{0,j}^{new};
7          if temp > max_val then
8              max_val = temp;
9              O_i = I_j;
10         end
11     end
12 end
```

5. Experiments

We conduct experiments on both of synthetic and real datasets to evaluate the proposed method. For comparison, we also compare the performance of our method with two other benchmark algorithms, including *majority voting*(MV) and *follow the best doctor*(FTBD). Where MV is a straightforward approach, which uses the most common label as the true label. From reported experimental results on real crowdsourcing data [10], MV performs significantly better on average than the individual workers. FTBD refers to a natural alternative for the patients that when they receive more than one diagnosis from different doctors, they will tend to follow the best doctors' diagnosis.

The detailed information of the real dataset is in Section 3, as to the constitution of the synthetic dataset, we adopt a 30×30 matrix R to represent the diagnosis results that are given by 30 doctors to 30 patients, where the rows correspond to the patients and columns to the doctors, hence, every entry $R_{i,j}$ is the diagnosis doctor D_j gives to patient P_i. We assume there are in total three illnesses I_1, I_2 and I_3, where every patient has equal probability to have one of the illnesses, so for each illness there are 10 patients with it. We observe that in real life every doctor often has some special diseases she has a good knowledge of, hence, for doctor $D_j (1 \leq j \leq 30)$ we draw a random number $x \sim Laplace(0, 1)$, where, when the patient has the $(1 + \lfloor \frac{j-1}{10} \rfloor) th$ illness, we assume D_j makes right the diagnosis with probability $1 - |x|$, and makes the wrong diagnosis to conclude that the patient has an arbitrary one of the other two illnesses with equal probability (*i.e.*, $\frac{1-|x|}{2}$).

We summarize the prediction performance in Tables 2–5, where the results on the synthetic data are presented in Tables 2 and 3, and the results on the real data are in Tables 4 and 5.

Table 2. The prediction accuracy on the synthetic data, higher is better.

Method	Accuracy (%)
MV	12(40.0%)
FTBD	20(66.7%)
CROWD	21(70%)

Table 3. The confusion matrix of the prediction on the synthetic data.

	MV			FTBD			CROWD		
	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
I_1	0.80	0.00	0.20	1.00	0.00	0.00	0.90	0.10	0.00
I_2	0.60	0.00	0.40	0.00	1.00	0.00	0.50	0.30	0.20
I_3	0.50	0.10	0.40	0.30	0.70	0.00	0.10	0.00	0.90

Table 4. The prediction accuracy on the real data, higher is better.

Method	Accuracy (%)
MV	24(48%)
FTBD	25(50%)
CROWD	28(56%)

Table 5. The confusion matrix of the prediction on the real data.

	MV				**FTBD**				**CROWD**			
	I_1	I_2	I_3	I_4	I_1	I_2	I_3	I_4	I_1	I_2	I_3	I_4
I_1	0.58	0.25	0.08	0.08	0.58	0.33	0.08	0.00	0.67	0.17	0.08	0.08
I_2	0.00	0.67	0.00	0.33	0.17	0.67	0.00	0.17	0.00	0.75	0.00	0.25
I_3	0.08	0.67	0.08	0.17	0.08	0.67	0.25	0.00	0.08	0.67	0.17	0.08
I_4	0.14	0.21	0.07	0.57	0.12	0.36	0.00	0.50	0.07	0.21	0.07	0.64

Tables 2 and 4 are for the prediction accuracy results, where we see our proposed algorithm outperforms the comparison methods on the both datasets. Tables 2 and 4 are the summarization of the confusion matrix of the results, where, for every algorithm, the $(i, j)th$ entry corresponds to the percentage value of the patients who are of illness i and diagnosed to be with illness j. For example, in Table 3, the top left entry $(I_1, I_1) = 0.80$ indicates that 80% of the I_1 patients are diagnosed correctly by the MV algorithm.

Another issue remained to be discussed is to address the value of λ. In our experiments, for the ith patient P_i, we calculate their λ value as follows:

$$\lambda = \frac{Number \quad of \quad diagnosis \quad to \quad the \quad all \quad patients}{Number \quad of \quad diagnosis \quad to \quad P_i} \tag{22}$$

Our intuitation of the definition is as follow: Denote Σ_1 as the number of diagnosis to the all patients, Σ_2 as the Number of diagnosis to P_i, it's clear that $\Sigma_1 \gg \Sigma_2$, so after the introduction of O_i, the divergence $KL(\mathcal{D}_i||\mathcal{D}_i^{new})$ is always far greater than $KL(\mathcal{D}_0||\mathcal{D}_0^{new})$, for compensation, we define λ as above.

6. Conclusions

Addressing the high level medical experts shortage problem, we present a crowdsourcing based scheme. Unlike the popular expert consultation systems, our approach aims at exploiting the power of the general practitioners' efforts. We propose a multiple diagnosis results ensemble policy, which is based on the statistical consistency w.r.t. the distribution of the results. We evaluate the proposed method on both the synthetic and real datasets. Results show it outperforms the comparison algorithms.

It is noteworthy that, although our algorithm yields better performance than the benchmarks in the empirical studies, and even the accuracy on the synthetic data is acceptable in practice, the results on the real data still remain not as high as expected. We think a main reason for this should be attributed to the limitation of the training data, because, in our experiment, all the volunteers are from the same department of the same medical school. Therefore, because of the reflection of their academic background, the diversity of their diagnosis results will be smaller than that of the real situation, or, in other words, the diagnosis results of different volunteers tends to be identical to each other. So when one volunteer has misdiagnosed a patient, it is most likely that many other volunteers will make the same mistake on the same patient, too. As a result, in the extreme case, no matter what the ensemble policy is, it is simply identical to the superposition of multiple duplicates. So, in our subsequent work, on the one side, we will try to introduce some small sample statistical technologies to improve the performance

of the algorithm, on the other side, we will keep on collecting more real data from different sources to enlarge the ground truth base.

Acknowledgments

This work is supported in part by Research Fund for the Doctoral Program of Higher Education of China (20120171120086), Educational Commission of Guangdong Province (2013113) and Science and Technology Planning Project of Guangdong Province (2012B061700078). The authors would like to thank Wubin Li for polishing the presentation.

Author Contributions

Jian-Yong Yang directed the research. Xian-Hong Xiang and Xiao-Yu Huang contributed equally in data analysis, algorithm design and paper writing. Xiao-Ling Zhang, Chun-Fang Cai and Lei Li helped to recruit the volunteers, collect the data and evaluate the model. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Roobottom, C.; Mitchell, G.; Morgan-Hughes, G. Radiation-reduction Strategies in Cardiac Computed Tomographic Angiography. *Clin. Radiol.* **2010**, *65*, 859–867.
2. Warach, S.; Gaa, J.; Siewert, B.; Wielopolski, P.; Edelman, R.R. Acute Human Stroke Studied by Whole Brain Echo Planar Diffusion-weighted Magnetic Resonance Imaging. *Ann. Neurol.* **1995**, *37*, 231–241.
3. Behrens, S.; Laue, H.; Althaus, M.; BÃűhler, T.; Kuemmerlen, B.; Hahn, H.K.; Peitgen, H.O. Computer Assistance for MR Based Diagnosis of Breast Cancer: Present and Future Challenges. *Comput. Med. Imaging Graph.* **2007**, *31*, 236–247.
4. Beijing. Available online: http://en.wikipedia.org/wiki/Beijing (accessed on 8 March 2014).
5. Shanghai. Available online: http://en.wikipedia.org/wiki/Shanghai (accessed on 8 March 2014).
6. Guangzhou. Available online: http://en.wikipedia.org/wiki/Guangzhou (accessed on 8 March 2014).
7. Muhammadi, J.; Rabiee, H.R. Crowd computing: A survey. **2013**, arXiv:1301.2774.
8. Howe, J. The rise of crowdsourcing. *Wired Mag.* **2006**, *14*, 1–4.
9. Howe, J. *Crowdsourcing: How the Power of the Crowd Is Driving the Future of Business*; Random House: New York, NY, USA, 2008.
10. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 254–263.

11. Bennett, J.; Lanning, S. The Netflix Prize. In Proceedings of KDD Cup and Workshop, San Jose, CA, USA, 12 August 2007; Volume 2007, p. 35.
12. Von Ahn, L.; Maurer, B.; McMillen, C.; Abraham, D.; Blum, M. recaptcha: Human-based character recognition via web security measures. *Science* **2008**, *321*, 1465–1468.
13. Von Ahn, L.; Dabbish, L. Labeling images with a computer game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004; ACM: New York, NY, USA, 2004; pp. 319–326.
14. Welinder, P.; Perona, P. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 25–32.
15. King, A.J.; Gehl, R.W.; Grossman, D.; Jensen, J.D. Skin self-examinations and visual identification of atypical nevi:comparing individual and crowdsourcing approaches. *Cancer Epidemiol.* **2013**, *37*, 979–984.
16. Amazon Mechanical Turk. Available online: http://aws.amazon.com/mturk/ (accessed on 8 March 2014).
17. Lintott, C.J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M.J.; Nichol, R.C.; Szalay, A.; Andreescu, D.; *et al*. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astronom. Soc.* **2008**, *389*, 1179–1189.
18. Kanefsky, B.; Barlow, N.G.; Gulick, V.C. Can distributed volunteers accomplish massive data analysis tasks. In proceedings of Lunar and Planetary Science, Houston, TX, USA, 12–16 March 2001; p. 1272.
19. Parshotam, K. Crowd computing: A literature review and definition. In Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, East London, South Africa, 7–9 October 2013; ACM: New York, NY, USA, 2013; pp. 121–130.
20. De, A.; Mossel, E.; Neeman, J. Majority is stablest: Discrete and SoS. In Proceedings of the 45th Annual ACM Symposium on Symposium on Theory of Computing, Palo Alto, CA, USA, 2013; ACM: New York, NY, USA, 2013; pp. 477–486.
21. Dekel, O.; Shamir, O. Good learners for evil teachers. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; ACM: New York, NY, USA, 2009; pp. 233–240.
22. Chen, S.; Zhang, J.; Chen, G.; Zhang, C. What if the irresponsible teachers are dominating? A method of training on samples and clustering on teachers. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010; pp. 419–424.
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 2012.