Application Of Combinatorial And Bayes' Theorem To Evaluate The Accuracy Of The Diagnosis Of Genetic Disorder

M. Aditya Farizki - 13516082
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13516082@std.stei.itb.ac.id

***Abstract*—There have been many cases of false diagnosis that causes severe effect to the patient. The use of advanced technology in the process have helped much but still we're far from perfect diagnosis even though the accuracy of the method is close to 100%. This paper explore the possibility that there's a fundamental problem with not being able to have 100% diagnosis accuracy to the result of the diagnosis itself, especially in rare genetic diseases using combinatorial and Bayesian theorem approach.**

***Keywords*—combination, probability, Mendel segregation law, genetic.**

I. Introduction

Diagnosis is the first step of treating a patient. Which also happen to be the most crucial part, because there are a lot of diseases with common symptoms need different way of treatment, just like in the case of avian flu and common flu. Things get even worse when it comes to genetic disease. The failure to diagnose a genetic disease often can be lethal.

The problem with genetic diseases is they are rare and hard to diagnose. One of the most common and reliable method to diagnose a disease is by using Differential Diagnosis technique, and the main problem is a genetic disease can and often has very similar symptoms with parasitic disease or more common disease, for example a common heart attack and Brugada syndrome almost cannot be differentiated.

It is not impossible to detect the presence of these abnormal genes in a human body with current technology. The computerization of diagnosing process can help a great deal when it comes to genetic disease as computer can collect and interpret the given symptoms much faster than human can ever do. In fact IBM is currently developing an AI for that purpose named IBM Watson.

With such technology we can gather data from around the world and make an incredibly accurate diagnosis. Using the same technique as what human doctor use, that is Differential Diagnosis but with a lot more data and faster thinking process.

This achievement is remarkable indeed, but this paper will explore the fundamental problem with not having the ability of 100% accurate diagnosis.

II. Combination, Bayes' Theorem, And Mendelian Inheritance

A. Combination

Combination is a special form of permutation in which permutation is part of the branch of mathematic called Combinatorial. The main idea of combination is to count how many possibilities of different selection given a set of object without regard to the order in which the objects are selected.

A classic example of combination application is in the ball in the box problem. Suppose that you have 2 identical balls which have the same color and 3 identical boxes. Every box can only contain 1 ball. How many ways to put the balls inside the box?

We can try to manually count the possible ways to put the ball. First we can put the first ball inside the first box then the second ball into the second box, now we can try to put the second ball into the first box and the first ball into the second box, but it will just be the same as the first step, the balls are identical and there's no way of telling which ball is which, so from these 2 steps we got 1 possibility. Next we can put the first ball in the second box and the second ball in the third box, this will be the same as putting the first ball in the third box and the second ball in the second box, hence we got 2 possible ways. Now we can put the first ball in the first box and the second ball in the third box, then we have 3 ways. Now we don't have any other way to put the ball in the box anymore, all possible ways have been done, so we end up with 3 ways to put two identical balls inside 3 identical boxes.

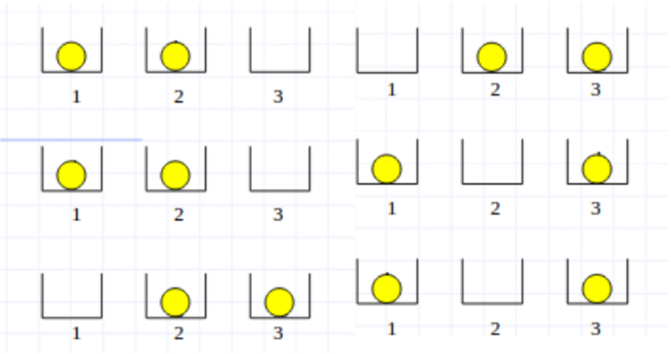

Illustration 1: solution to the ball problem, cited from http://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2014-2015/Kombinatorial%20(2014).pdf accessed on 30 November 2017

We can formulate the total amount of possible ways to put the ball. If we rephrase the question to how many possible ways to chose which box to be filled with the ball? Now it seems clearer that the ball problem can be interpreted as a permutation problem. We'd have 3 ways to chose to which box will be filled with the first ball, and we'd have 2 ways to chose to which box will be filled with the second ball, so in the end we'd have a total of 6 ways to chose which box to be filled with which ball as shown in the illustration 1. But we have duplicates so we must rule out the duplicate and leaving us we 3 ways. Then the formula for this problem can be written in permutation form.

$$\frac{P(3,2)}{2} = \frac{P(3,2)}{2!} = 3$$

We can generalize the used formula to a form that can solve all similar problem. The idea is the total amount of all possible outcome for a combination of matching n things to r room is the factorial of n divided by the factorial of r times the factorial of n – r.

$$C(n,r)=\frac{n!}{r!(n-r!)}$$

B. Bayes' Theorem

The idea behind Bayes Theorem is that if you have more information then you can get a more realistically accurate conclusion from the data. Bayes Theorem has arguably counter-intuitive nature, because from a high probability that an event occur given a hypothesis is true, you can get a low probability that the hypothesis is true given the event occur. Bayes Theorem is formulated like this :

$$P(H|E)=\frac{P(E|H)*P(H)}{P(E)}$$

H = Hypothesis

E = Event

P(H|E) = the probability of the hypothesis is true given an event occur

P(E|H) = the probability that an event occur given the hypothesis is true

P(H) = the prior probability that the hypothesis is true

P(E) = the probability that the event occurring.

The probability of a hypothesis given an event is true is equal to the probability that an event occur given the hypothesis is true times the prior probability that a hypothesis is true divided by the probability that the event occurring.

The prior probability of the hypothesis being true is most of the time the hardest part of the equation to figure out. A example on how you determine the prior probability of a hypothesis is true is by using data of occurrence in a certain frequencies, for instance like how often an email is categorized as a junkmail for every 1000 emails.

The most interesting of Bayes Theorem is how to probability of a hypothesis is true given an event occur increase rapidly as the number of samples increase. For an extreme case such as rare genetic disorder the increase can be up to 10 times for only double the number of samples. This is also the most important part of Bayes Theorem where it states that more data can give you tremendously higher accuracy in probability.

C. Mendelian Inheritance

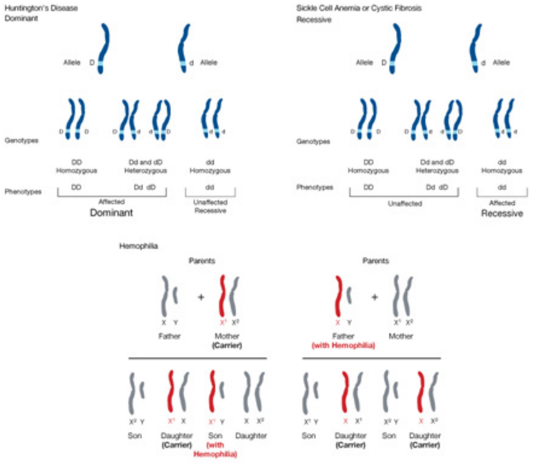

Illustration 2: the inheritance of hemophilia disease, cited from https://geneed.nlm.nih.gov/images/mendelian_inheritance_sm.jpg accessed on Friday 1st December 2017

1. Mendel's Law Of Segregation

Mendel's Law Of Segregation states that for each traits in an individual organism there are two alleles that define the trait, and these alleles segregate during meiosis such that each gamete contains only on of the alleles. Then the offspring of the individual will receive a pair of alleles from the mother and father, and inherit homologous chromosomes for a trait, one allele from each parent. During gamete formation the two alleles of the parents will separate and each gamete will have only one allele for each trait.

2. Mendel's Law Of Independent Assortment

The Law Of Independent Assortment states that for each alleles that carry different traits are independent of each other. So the fact that allele A is passed for trait Z doesn't affect which allele is passed for trait X.

3. Mendel's Law Of Dominance

Mendel's Law Of Dominance states that a dominant allele will always cover/mask the recessive alleles. Therefore a cross between recessive allele and dominant allele will always results in the trait of the dominant allele

4. Mendelian Trait

Mendelian trait is a trait that is controlled by a single locus in an inheritance pattern, which means the trait is directly controlled by which allele carries the certain trait. It also implies that some certain diseases and disorder can be affected directly from the allele, examples include sickle-cell anemia, Thalassemia and cystic fibrosis.

Those are the four main points of Mendelian Inheritance, for the scope of this paper we will cover only the disorder or disease that's fulfill the definition of Mendelian Trait.

There are also a type of inheritance that to a certain degree defies the Mendelian inheritance, for example a cross between two *Mirabilis jalapa* plants shows that the first generation of the offspring have an appearance in between the two parents. Some alleles are neither dominant nor recessive. In this paper for the sake of clarity and simplicity, we will ignore the existence of such inheritance. Because it will make the inheritance harder to recognize and isolate to which gene is responsible for which trait.

III. Abstract Case Analysis

The purpose of abstract case analysis is to give you an idea about the probability of getting the right diagnosis given a certain extreme case condition. The given number and condition is not entirely made up, it's based on a rare genetic disorder which is hemophilia, but we skip some trivial detail for the sake of simplicity, so that the reader may understand the consequence of the application of the given theorem on the subject.

Given a population of one thousand people. Let's say that there's a disease, a horrible and extremely rare disease in the population with the probability of 0.1% of the population having the disease, if we have a set of apparatus in which can identify the disease with 99% accuracy, what is the probability that a person who's diagnosed with the disease actually have the disease?

We can use Bayes' Theorem directly for this problem. The probability of the person is actually sick and is diagnosed sick will be represented by P(H|E), the accuracy of the test apparatus will be represented by P(E|H), the P(E|H) actually means the probability that the event (the person is diagnosed as sick) occur and the the hypothesis of the person is actually sick is true, which is equivalent with the accuracy of the apparatus in diagnosing a person sickness. The P(H) will be the probability of a person having a disease in the population which is given by the data. For the P(E), probability of a person is diagnosed as sick, is equal to the probability of the person is diagnosed sick and actually is sick, plus the probability of the person is diagnosed as sick but actually he's not sick. The equation end up looking like this :

$$P(H|E)=\frac{P(E|H)*P(H)}{P(E|H)*P(H)+P(\neg H)*P(E|\neg H)}$$

P(E|H)	= 0.95
P(H)	= 0.001
P(-H)	= 0.999
P(E|-H)	= 0.05

plugging in the numbers and we get the result of the equation.

$$P(H|E)=\frac{0.95*0.001}{0.95*0.001+0.999*0.05}$$

$$P(H|E)\approx 0.02$$

We get a result of 2% accuracy of a person who is diagnosed as sick and is actually sick despite having an apparatus that has the accuracy of 95%.

This is the counter-intuitive nature of Bayes' Theorem, we have an apparatus that's so precise yet the accuracy of our diagnosis is only up to 2%, which mean out of 50 people who are diagnosed to have the disease only 1 of them actually has the disease.

But actually we can try to make sense with this result if we try to see the problem in a different approach. Let's imagine a population of one thousand people, with 0.1% probability of someone is actually sick then we know there's only 1 person in that population which is actually sick. Now with 95% accuracy we know that if everyone is tested with the apparatus then another 50 people will be diagnosed as sick too, because the apparatus fail for 5% of the test. Which means there are 51 people are diagnosed as sick but only 1 people is actually sick, hence we get the accuracy of 1/51 which is approximately 2%.

This result is so bizarre and frightening, it implies that a really rare disease is extremely hard to diagnose and no matter how well the apparatus at diagnosing the disease as long as it doesn't have 100% accuracy then the probability that you get the wrong diagnosis result is very high.

However the probability of false diagnosis can be reduced significantly if we try to have a second opinion, or at least get tested somewhere else again.

Let's say that all of the 51 people get diagnosed again by different doctor in different location and different lab, now we can apply the Bayes' Theorem again. The probability of the person is actually sick and is diagnosed sick will be represented by P(H|E), the accuracy of the test apparatus will be represented by P(E|H), in this case the accuracy of the apparatus is equal to the accuracy of the apparatus from the previous diagnosis, the P(E|H) actually means the probability that the event (the person is diagnosed as sick) occur and the the hypothesis of the person is actually sick is true, which is equivalent with the accuracy of the apparatus in diagnosing a person sickness. The P(H) will be the key part, in the previous diagnosis the P(H) is equal to the probability of a person having the disease in the population, but in the second diagnosis the P(H) is the probability of the person in the group of 51 people who take the diagnosis again is actually sick, this information gets updated each time we redo the diagnosis and drastically increase the accuracy of the result. For the P(E), probability of a person is diagnosed as sick, is equal to the probability of the person is diagnosed sick and actually is sick, plus the probability of the person is diagnosed as sick but actually he’s not sick. After plugging the new number we get the new result which is :

$$P(H|E)=\frac{0.95*0.02}{0.95*0.02+0.05*0.98}$$

$$P(H|E)\approx 0.28$$

By redoing the test we increase the accuracy by more than 10 times. This shows the the fundamental property of Bayes’ Theorem where when given more data, then the accuracy of the result will increase drastically. Now let’s see how accurate we can get if we redo the test once more time, of course with the assumption that the test is carried in a different lab and different apparatus but with the same precision.

$$P(H|E)=\frac{0.95\times 0.28}{0.95\times 0.28+0.05\times 0.72}$$

$$P(H|E)\approx 0.88$$

At the third test we get the accuracy of 88% which is really good considering our initial result’s accuracy is 2%. The introduction of new data will keep increasing the accuracy but will never reach 100% as long as our apparatus is not capable of having 100% accuracy, therefore the accuracy of our result is asymptotic to 100% as we get more and more data.

IV. Realistic Case Analysis

For the realistic case analysis we will be focusing on one genetic disorder which is thalassemia. The reason that thalassemia is picked to be our main focus is because the disorder is rare, the explanation for the inheritance of the disease is relatively easy to understand and doesn’t require too much data, thalassemia is a Mendelian trait, and the accuracy of the method to diagnose thalassemia can be categorized as low.

Thalassemia is an abnormality in alpha globin gene or beta globin gene, Thalassemia can be categorized into two type, alpha thalassemia and beta thalassemia, in this paper we will not differentiate the two, we will refer them only as thalassemia, this is done because the mutation and inheritance is identical and doesn’t affect the result of the calculation.

1. Thalassemia Inheritance

let’s name the alpha globin and beta globin allele that cause thalassemia as ‘s’, capitalized ‘s’ means dominant allele which mean a healthy normal allele, and uncapitalized ‘s’ as the recessive allele which mean it carries the thalassemia disease. The possible combination of the trait are SS, Ss and ss. SS will be a healthy normal person, Ss is a carrier, which mean he’s not sick but he has the potential to inherit the sickness and ss is the person with thalassemia.

We can calculate the amount of all possible combination from the given allele and we can get the probability of the offspring carrying or having the disease.

The amount of all possible offspring outcome given the parents have the most variance of the allele is the combination of choosing one allele from an ‘Ss’ mother and ‘Ss’ mother.

$$N=C(2,1)\times C(2,1)$$

$$N=4$$

we now can simulate the probability of the offspring of each possible pair of allele.

Table 1: Ss x Ss

SS	Ss
Ss	ss

Table 2: SS x Ss

SS	Ss
SS	Ss

Table 3: SS x ss

Ss	Ss
Ss	Ss

Table 4: Ss x ss

Ss	ss
ss	ss

Table 5: ss x ss

ss	ss
ss	ss

Table 6: SS x SS

SS	SS
SS	SS

2. Population Construction

We can now simulate the spread of thalassemia from the first time the disease is detected. We can assume that at the first mutation, the gene did not get corrupted at both allele so the first occurrence of the disease was caused by both parents being a carrier, after that assuming that there’s no any migration in the population we can simulate the spread of thalassemia.

● = person with no dominant allele (sick) (ss)

○ = healthy person (SS)

🔵 = carrier (Ss)

Note : the graph is created with the consideration that it can demonstrate the spread of the disease via inheritance, by no means I, the writer, support any form of incest relationship nor something along those lines.

We will assume that the healthy one will always marry the sick one, and the sick one will have offspring before dying.

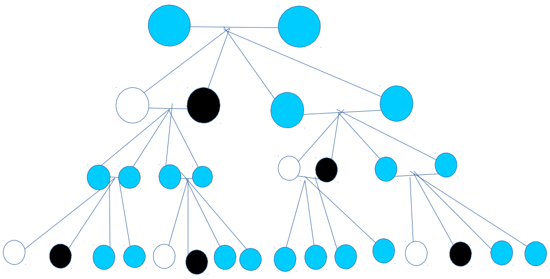

Illustration 3 : the inheritance of thalassemia via mating

The graph is basically a tree, by the tenth generation there will be around one thousand people in the population from the two parents with carrier allele. It can be counted the equation as follows (we consider the first parent as the first generation):

$$N=(branch/mate)^{generation}$$

N is the number of the population of the nth generation, branch/mate is the number of branch or offspring for each parents pair divided by the number of individuals needed to continue the regeneration. We know that each time a parent pair mate they will have 4 offspring and we need 2 individuals to continue the regeneration so the value of branch/mate in this case is 2. if we want to have a population of at least one thousand people, then we must be at the 10th generation at least. The ratio of sick people in a generation is approaching the number of approximately 17%, this can be proven with statistical evidence.

Table 7: sick people percentage

population	sick population	percentage
4	1	0.25
8	1	0.125
16	3	0.1875
32	5	0.15625
64	11	0.171875
128	21	0.1640625
256	43	0.16796875
512	85	0.166015625
1024	173	0.1689453125

3. Bayes' Theorem Application

The accuracy of our current technology to detect thalassemia is at around 89%. With these data we can apply the Bayes' Theorem to count the probability of someone actually having thalassemia given the diagnosis result is positive.

Applying the numbers to the Bayes' equation will give us the following result :

$$P(H|E)=\frac{0.89\times0.17}{0.89\times0.17+0.83\times0.11}$$

$$P(H|E)\approx0.62$$

We got the accuracy of someone actually is sick given the diagnosis result is positive is 62, which is relatively high compared to the abstract case analysis, but still it means out of one hundred people, 48 people will have false diagnosis.

If we redo the test like in the abstract case analysis we'll get the following result :

$$P(H|E)=\frac{0.89\times0.62}{0.89\times0.62+0.38\times0.11}$$

$$P(H|E)\approx0.93$$

by getting a second opinion we'll increase the probability that the sick person is diagnosed as sick up to 93%, which is to a certain degree, can be considered safe.

Something to be noted is that we have been pretty generous with the number, thalassemia's occurrence is not as frequent as the occurrence in the simulated population, but still, with the given data we can simulate the accuracy of diagnosing a genetic disorder to a point that's close to reality.

This part shows a more realistic and more common case in diagnosing an illness. The abstract case analysis part gives us the idea of how bad the problem is in extreme cases. But there are so many other and rarer disease than thalassemia in the world. Diseases such as the Dawson disease or even something that was common like the Bubonic plague or even rabies. It is important for us to keep developing our technology so that we can have more and more accurate apparatus and minimize the problem that's addressed in this paper.

In real word, real doctor and medical institution will give more effort and repeated test in order to accurately diagnose a patient sickness, but that does not mean the content of this paper is irrelevant, the probability of a patient getting a false diagnosis is still high if the probability of someone getting the disease in the population is low enough, some of theses cases are brugada syndrome, bubonic plague, als disease and many more, it's important to get a second opinion or even third opinion when you're dealing or having the specific symptoms of those diseases.

V. Conclusion

With the simulated environment we get the result of at most 62% of accuracy when diagnosing a patient with thalassemia, this is considerably high but still not safe enough, by getting a second opinion the probability that the patient is actually have thalassemia and is diagnosed with thalassemia increased to 93% which is why second opinion is a very important part if you're dealing with rare disease. Lastly in the real world, almost all medical institution will have more measures in which they will increase the accuracy of their diagnosis, and the measurement on this paper is not to be taken without consideration, but the point is the liability is there, there's a mathematical proof that when a disease is rare enough, the step or effort to diagnose a patient with the disease accurately is incredibly high.

VI. Acknowledgment

The author would like to thank Dr. Ir. Rinaldi Munir, MT. as the lecturer of Discrete Mathematic class, for guidance in preparing this paper. The author would also like to thank the various researchers who post their journal and make it easily accessible to the author.

References

[1] K.H. Rosen, Discrete Mathematics and its Applications, 7th ed. New York: McGraw-Hill, 2012, pp. 641802.

[2] https://www.ncbi.nlm.nih.gov/pubmed/21774279
(last accessed in 2nd December 2017 23.43)

[3] http://thalassemia.com/what-is-thal-alpha.aspx#gsc.tab=0
(last accessed in 3rd December 2017 12.25)

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 3 Desember 2017

M. Aditya Farizki-13516082