Granularity Explains Empirical Factor-of-Three Relation Between Probabilities of Pulmonary Embolism in Different Patient Categories

Beverly Rivera[1], Francisco Zapata[2], and
Vladik Kreinovich[1,2]
[1]Computational Science Program
[2]Center for Theoretical Research and its
Applications in Computer Science (TRACS)
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
barivera@miners.utep.edu, fazg74@gmail.com,
vladik@utep.edu

Abstract

Pulmonary embolism is a very dangerous difficult-to-detect medical condition. To diagnose pulmonary embolism, medical practitioners combine indirect signs of this condition into a single score, and then classify patients into low-probability, intermediate-probability, and high-probability categories. Empirical analysis shows that, when we move from each category to the next one, the probability of pulmonary embolism increases by a factor of three. In this paper, we provide a theoretical explanation for this empirical relation between probabilities.

1 Formulation of the Problem

Pulmonary embolism: a brief reminder. One of the most dangerous medical conditions is *pulmonary embolism*, a blockage of the main artery of the lung (or one of its branches) which can lead to collapse and sudden death; see, e.g. [1]. Pulmonary embolism is responsible for about 15% of sudden deaths.

If detected on time, pulmonary embolism can be treated: either by anticoagulation medicine like heparin or warfarin, or – in severe cases – by a surgery. The problem is that pulmonary embolism is difficult to diagnose: lungs are mostly normal, fever is either absent or low-grade, etc.

Scores: a brief description. Since pulmonary embolism is difficult to directly diagnose, hospitals' emergency departments take into account different variables

like age, heart rate, different types of pain, etc., to produce a numerical score. A high score indicates a high probability of pulmonary embolism; so, the doctors start applying aggressive treatment to such patients.

One of the most widely used ways to assign scores is known as the *Geneva score*; its latest version is described by [4]. Depending on this score, patients are classified into three categories:

- patients with low scores are classified into the *low-probability* category;
- patients with intermediate scores are classified into the *intermediate-probability* category; and
- patients with high scores are classified into the *high-probability* category.

Scores: empirical fact. According to an empirical study [4]:

- in the low-probability category, approximately 8% of the patients had pulmonary embolism;
- in the intermediate-probability category, approximately 28% of the patients had pulmonary embolism; and
- in the high-probability category, approximately 74% of the patients had pulmonary embolism.

From each category to the next one, the probability increases by a factor of three.

What we do in this paper. Division into categories is a particular case of *granularity*; see, e.g., [5]. In this paper, following ideas from [2, 3], we use granularity techniques to provide a theoretical explanation for the above empirical relation between probabilities.

2 Explanation

Main idea. We are interested in the situation where we estimate probability – a quantity which can only take non-negative values. In general, to estimate the values of a non-negative quantity, we select a sequence of positive numbers $\ldots < e_0 < e_1 < e_2 < \ldots$ (e.g., 0.1, 0.3, 1.0, etc.), and every actual value x of the estimated quantity is then estimated by one of these numbers. Each estimate is approximate: when the estimate is equal to e_i, the actual value x of the estimated quantity may differ from e_i; in other words, there may be an estimation error $\Delta x = e_i - x \neq 0$.

What is the probability distribution of this estimation error? This error is caused by many different factors. It is known that under certain reasonable conditions, an error caused by many different factors is distributed according to Gaussian (normal) distribution (see, e.g., [7]; this fact – called *central limit theorem* – is one of the reasons for the widespread use of Gaussian distribution

in science and engineering applications). It is therefore reasonable to assume that Δx is normally distributed.

It is known that a normal distribution is uniquely determined by its two parameters: its average a and its standard deviation σ. Let us denote the average of the error Δx by Δe_i, and its standard deviation by σ_i. Thus, when the estimate is e_i, the actual value $x = e_i - \Delta x$ is distributed according to Gaussian distribution, with an average $e_i - \Delta e_i$ (which we will denote by $\widetilde{e}_i$), and the standard deviation σ_i.

For a Gaussian distribution with given a and σ, the probability density is everywhere positive, so theoretically, we can have values which are as far away from the average a as possible. In practice, however, the probabilities of large deviations from a are so small that the possibility of such deviations can be safely neglected. For example, it is known that the probability of having the value outside the "three sigma" interval $[a-3\sigma, a+3\sigma]$ is $\approx 0.1\%$ and therefore, in most applications to science and engineering, it is assumed that values outside this interval are impossible.

There are some applications where we cannot make this assumption. For example, in designing computer chips, when we have millions of elements on the chip, allowing 0.1% of these elements to malfunction would mean that at any given time, thousands of elements malfunction and thus, the chip would malfunction as well. For such critical applications, we want the probability of deviation to be much smaller than 0.1%, e.g., $\leq 10^{-8}$. Such small probabilities (which practically exclude any possibility of an error) can be guaranteed if we use a "six sigma" interval $[a - 6\sigma, a + 6\sigma]$. For this interval, the probability for a normally distributed variable to be outside it is indeed $\approx 10^{-8}$.

Within this Gaussian description, what is the optimal granularity?

Optimal granularity: informal explanation. In accordance with the above idea, for each e_i, if the actual value x is within the "three sigma" range $I_i = [\widetilde{e}_i - 3\sigma_i, \widetilde{e}_i + 3\sigma_i]$, then it is reasonable to take e_i as the corresponding estimate.

We want a granulation which would cover all possible values, so each positive real number must be covered by one of these intervals. In other words, we want the union of all these intervals to coincide with the set of all positive real numbers.

We also want to makes sure that all values that we are covering are indeed non-negative, i.e., that for every i, even the extended "six sigma" interval

$$[\widetilde{e}_i - 6\sigma_i, \widetilde{e}_i + 6\sigma_i]$$

only contains non-negative values.

One of the main purposes of granularity is to decrease the number of "labels" that we use to describe different quantities. So, we want to consider optimal (minimal) sets of intervals. Formally, we can interpret "minimal" in the sense that whichever finite subset we pick, we cannot enlarge their overall coverage by modifying one or several of these intervals. Let us formalize these ideas.

Optimal granularity: formal description. In the following definitions, we will use the fact that an arbitrary interval $[a^-, a^+]$ can be represented in the

Gaussian-type form $[a-3\sigma, a+3\sigma]$: it is sufficient to take $a = (a^- + a^+)/2$ and $\sigma = (a^+ - a^-)/6$.

Definition.

- *We say that an interval $I = [a-3\sigma, a+3\sigma]$ is* reliably non-negative *if every real number from the interval $[a-6\sigma, a+6\sigma]$ is non-negative.*
- *A set $\{I_i\}$, $i = 1, 2, \ldots$, of reliably non-negative intervals I_i is called a* granulation *if every positive real number belongs to one of the intervals I_i.*
- *We say that a granulation* can be improved *if, for some finite set $\{i_1, \ldots, i_k\}$, we can replace intervals I_{i_j} with some other intervals I'_{i_j} for which*
$$\bigcup_{j=1}^{k} I_{i_j} \subset \bigcup_{j=1}^{k} I'_{i_j} \quad \bigcup_{j=1}^{k} I_{i_j} \neq \bigcup_{j=1}^{k} I'_{i_j},$$
and still get a granulation.
- *A granulation is called* optimal *if it cannot be improved.*

Proposition. *In an optimal granulation, $I_i = [a_i, a_{i+1}]$, where $a_{i+1} = 3a_i$.*

This explains the fact that each next probability is three times larger than the previous one.

Proof.

1°. Let us first prove that for every interval $I_i = [a_i - 3\sigma_i, a_i + 3\sigma_i]$ from an optimal granulation, $a_i = 6\sigma_i$.

Indeed, since all the intervals I_i must be reliably non-negative, we can conclude that $a_i - 6\sigma_i \geq 0$, hence $a_i \geq 6\sigma_i$. So, to complete this part of the proof, it is sufficient to show that we cannot have $a_i > 6\sigma_i$. We will prove this by showing that if $a_i > 6\sigma_i$, then the corresponding granulation can be improved.

Indeed, in this case, we can take $\sigma'_i = a_i/6 > \sigma_i$, and consider a wider interval $I'_i = [a_i - 3\sigma'_i, a_i + 3\sigma'_i] \supset I_i$. Due to our choice of σ'_i, this new interval is also reliably non-negative. Therefore, if we replace the interval I_i by I'_i, we still get a granulation, and $I_i \subset I'_i$, $I_i \neq I'_i$. Thus, the original granulation can be improved.

So, if the granulation is optimal (i.e., cannot be improved), we have $a_i = 6\sigma_i$.

2°. Let us now prove that for every interval $I_i = [a_i^-, a_i^+]$ from an optimal granulation, $a_i^+ = 3a_i^-$.

Indeed, from Part 1 of this proof, we can conclude that for an arbitrary interval $I_i = [a_i^-, a_i^+] = [a_i - 3\sigma_i, a_i + 3\sigma_i]$ from the optimal granulation, we have $3\sigma_i = 0.5 \cdot a_i$, hence $a_i^- = a_i - 3\sigma_i = 0.5 \cdot a_i$ and $a_i^+ = a_i + 3\sigma_i = 1.5 \cdot a_i$. Thus, $a_i^+ = 3a_i^-$.

3°. Let us now show that if two intervals from an optimal granulation intersect, then this intersection can only consist of a single point.

To prove this, we will show that if two intervals $I_i = [a_i^-, a_i^+]$ and $I_j = [a_j^-, a_j^+]$ have a more extensive intersection, then the granulation can be improved. Without losing generality, we can assume that $a_i^- \le a_j^-$.

We already know that since both I_i and I_j are intervals from an optimal granulation, we have $a_i^+ = 3a_i^-$ and $a_j^+ = 3a_j^-$. Since $a^- \le a_j^-$, we thus conclude that $a_i^+ = 3a_i^- \le 3a_j^- = a_j^+$.

The fact that the intervals $I_i = [a_i^-, 3a_i^-]$ and $I_j = [a_j^-, 3a_j^-]$ have an intersection means that $a_j^- \le 3a^-$; the fact that this intersection is not simply a single point means that $a_j^- < 3a_i^-$. In this case, $I_i \cup I_j = [a_i^-, 3a_j^-]$.

Let us show that we can improve the granulation if we replace I_i by itself $I_i' = I_i$ and I_j by $I_j' = [3a_i^-, 9a_i^-]$. Indeed, both new intervals are reliably non-negative, and the new union $I_i' \cup I_j' = [a_i^-, 9a_i^-]$ is a strict superset of the old one – because $a_j^- < 3a_i^-$ hence $3a_j^- < 9a_i^-$.

4°. So, in an optimal granulation, every interval must be of the type $[a, 3a]$, these intervals must cover the entire real axis, and they cannot intersect in more than one point. Thus, right after each interval $[a_i, 3a_i]$, there should be the next interval $[a_{i+1}, 3a_{i+1}]$, so we should have $a_{i+1} = 3a_i$.

Thus, we get the description from the formulation of the theorem.

5°. One can also easily prove that the granulation in which $I_i = [a_i, a_{i+1}]$ with $a_{i+1} = 3a_i$ cannot be improved and is thus optimal. The proposition is proven.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

References

[1] S. Z. Goldhaber, "Pulmonary thromboembolism", In: by D. L. Kasper, E. Braunwald, A. S. Fauci, S. L. Hauser, D. Longo, and J. L. Jameson (eds.), *Harrison's Principles of Internal Medicine*, McGraw Hill, Columbus, Ohio, 2004, pp. 1561–1565.

[2] J. R. Hobbs and V. Kreinovich, "Optimal choice of granularity in commonsense estimation: why half-orders of magnitude", *Proceedings of the Joint 9th World Congress of the International Fuzzy Systems Association and 20th International Conference of the North American Fuzzy Information Processing Society IFSA/NAFIPS 2001*, Vancouver, Canada, July 25–28, 2001, pp. 1343–1348.

[3] J. Hobbs and V. Kreinovich, "Optimal choice of granularity in commonsense estimation: why half-orders of magnitude", *International Journal of Intelligent Systems*, 2006, Vol. 21, No. 8, pp. 843–855.

[4] G. Le Gal, M. Righinin, P. M. Roy, O. Sanchez, D. Aujesky, H. Bouameaux, and A. Perrier, "Prediction of pulmonary embolism in the emergency department: the revised Geneva score", *Annals of Internal Medicine*, 2006, Vol. 144, No. 3, pp. 165–171.

[5] W. Pedrycz, A. Skowron, and V. Kreinovich (eds.), *Handbook on Granular Computing*, Wiley, Chichester, UK, 2008.

[6] B. Pregerson, *Quick Essentials: Emergency Medicine*, E.D. Insight Books, Bel Air, Califonia, 2004.

[7] H. M. Wadsworth, Jr. (editor), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., N.Y., 1990.