COMPARING G WITH CLASSICS PATTERN RECOGNITION FORMULAE IN AI

Author 1: *Claudio Garuti*
Author 2: Fabiola Garuti

Highlights

- The G index demonstrates superior performance over traditional AI methods in weighted environments, offering enhanced sensitivity and accuracy.
- This study introduces the Maximum Difference Formula, enabling precise identification of maximum deviation points in pattern recognition.
- Application of the G index improves diagnostic accuracy, making it a valuable tool for sensitive fields like medical diagnosis.

ABSTRACT

This study assesses the G index as a similarity measure for pattern recognition, particularly in weighted environments like medical diagnosis. Compared to traditional methods such as Weighted Absolute Differences (WAD) and Dot Product (DP), the G index provides greater sensitivity and accuracy using the definition of cosine function, normalizing each coordinate and sum over all the profile's coordinates. Through analysis and examples, the paper demonstrates that G better captures variations in profile similarity, especially with differing indicator weights. The study also introduces the Maximum Difference Formula for pinpointing maximum deviation, enhancing database optimization for disease profiles. This research highlights the importance of choosing appropriate similarity measures to improve diagnostic accuracy in sensitive fields through better pattern recognition.

Keywords: AI, Pattern recognition, G index, Weighted Absolute Differences, Dot Product, medical diagnosis, weighted environments

1. Introduction

The current rise of artificial intelligence has unleashed a series of chain reactions in all fields. It is no surprise that this technology has revolutionized the world as we know it today, increasing the efficiency of many processes and significantly supporting decision-making through its weighted and intelligent algorithms. However, it is important not to blindly trust this new technology. While it can be very useful in numerous areas, its lack of sensitivity in weighted environments can lead to fatal errors in others, such as in medical diagnosis. For this reason, there is a need to compare the well-known G index, a measure of similarity and compatibility, with the 3 main methods used by AI: Weighted Absolute Differences (Manhattan distance), Square root of weighted square differences (Euclidean distance) and the Vectors inner product (Dot Product), to determine which of these is better suited for sensitive cases such as medical diagnosis.

2. Literature Review

This study builds on foundational research in similarity measures and pattern recognition, particularly in the context of weighted environments. A primary influence is Saaty's Analytic Hierarchy Process (AHP), which provides a robust framework for weighted decision-making in complex scenarios (Saaty, 2005). While AHP effectively assigns weights to various criteria, it lacks a mechanism for direct similarity measurement between profiles, highlighting a gap that the G index seeks to address.

The limitations of traditional similarity measures, such as Dot Product (Cosine Similarity) and Weighted Absolute Differences (WAD), are well-documented in the literature. For instance, Singh and Gupta (2010) note that cosine similarity is limited in applications where variations in weight significantly impact accuracy. Similarly, Chen and Du (2008) discuss how WAD fails to account for weight sensitivity in clinical pattern recognition, often resulting in inaccuracies that can affect outcomes in sensitive fields like medical diagnostics. These critiques support the need for a more nuanced similarity measure that can accurately reflect weighted differences.

Further, Kononenko (2001) underscores the importance of precise similarity measures in medical diagnosis, where pattern recognition plays a critical role. His work highlights that while AI and machine learning techniques are advancing in the medical field, traditional methods may lack the necessary sensitivity, particularly in weighted environments.

Garuti's (2020) introduction of the G index provides a promising solution to these issues, demonstrating the index's potential in capturing nuanced variations in weighted environments. By normalizing compatibility on a scale using minimum and maximum values, the G index offers a more sensitive and accurate approach to similarity measurement than WAD or DP, particularly useful in high-stakes applications like medical diagnostics.

3. Hypotheses/Objectives.

The primary hypothesis is that the G index outperforms both weighted absolute differences (WAD) and dot product in weighted environments, providing greater sensitivity and accuracy into predicting pattern recognition into weighted environments. Specifically, the objective is to demonstrate that G allows more controlled sensitivity to changes in similarity, enhancing decision-making accuracy in weighted pattern recognition, particularly in medical diagnosis scenarios.

Based on this premise, the following specific objectives are proposed:

- Demonstrate through comparison that G is a better indicator than WAD, as the latter may drive to wrong choices when selecting the coordinate with the greater variation.
- Demonstrate through comparison that G is better than dot product, given that, although both show the same similarity trend, G represents a better change behavior (the sensitive is more controlled) and hence more accuracy in the final result.

4. Research Design/Methodology

In order to understand the foundations of the comparison, one must first grasp the difference between the two types of measurement: statistical and topological. The former is based on measuring frequency-based trends, which is useful for relating and structuring datasets in either an ordinal or cardinal manner. However, this type of measurement is clearly relative to the dataset, as statistical measurement depends on it. Similarly, normalization is dependent on the data, so the properties of the data may be altered, completely changing the results.

Regarding topological measurement, it involves measures of distance or proximity, which are useful for prioritizing patterns and their corresponding objectives in a cardinal manner. Unlike statistical measurement, topological measurement does not depend on the dataset, meaning that the acceptance/rejection threshold for similarity is absolute. This also implies that normalization is absolute, so it does not alter the properties of the data.

For medical diagnosis, the human body is topographically divided in 35 AHP models with a list of disease profiles that were in a large ANP model. Feedback between the AHP models and the symptoms and signs of the disease profiles appear, in which each disease activate a model's subset from the 35 initial models, conforming an Holarchy where the alternatives (the diseases) affect (active) the main criteria (the AHP models):

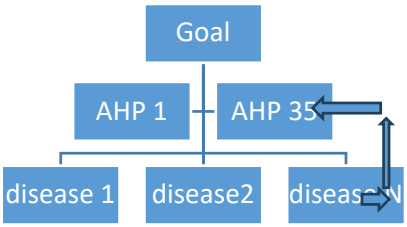


Figure 1: The ANP Holarchy

Every disease profile DP is a function of two parameters, Weight and Intensity of symptoms and Signs (S&S) which define that DP. Thus, DP = $f(W, I)$. Symptoms are divided in general (without a specific location in the body) and specific (associated to more specific DP and normally with a specific location). Every AHP model have general symptoms, then it may repeat and hence overrepresent those symptoms. Thus, after the feedback process it's necessary to "clean" the redundancies in general symptoms and link it with the specific symptoms of each model.

Let's explore the concept of proximity measurement in a medical context by comparing two approaches: absolute variation and relative variation for measuring symptom intensities and weights. The goal is to find the maximum variation between a "pattern" (standard or expected profile of symptoms) and a "user response" (actual patient symptom profile), focusing in improving the accuracy of medical judgments by identifying which approach captures variations more effectively. It´s important to notice the two main pieces of data required: symptom importance (w) and the symptom or sign intensity (I).

The formulas commonly used in AI and Multi-Criteria Decision Making (MCDM) for determining the maximum variation from the pattern consider symptom importance (w) and symptom intensity (y and x) are:

$$\boldsymbol{Max_i\{w * (y - x)\}} \quad \textbf{(1)}$$

$$\boldsymbol{Max_i\{\sqrt{w^2 * (y - x)^2}\}} \quad \textbf{(2)}$$

However, this paper proposes a variation in the form of:

$$\boldsymbol{Max_i\{w * (1 - x/y)\}} \quad \textbf{(3)}$$

Which incorporates a relative difference rather than an absolute one, with w, x, y between 0 and 1, $y > x$ and $i = 1, n°indicators$, i.e., symptoms.

In medicine, adequate pattern recognition leads to successful diagnosis and, therefore, higher probabilities of resolution and/or effective treatment. For this reason, it is important to be able to recognize patterns successfully by the comparison of two or more profiles (A and B in this case), defined as a set of weighted intensities for different coordinates (or symptoms) based on their similarity. Here, the objective is to evaluate how close these two profiles are and explore methods to precisely measure this compatibility. Mathematically:

$$\boldsymbol{ProfileA} = \sum_i(\boldsymbol{w_i * Ia_i}) \quad \textbf{(4)}$$

$$\boldsymbol{ProfileB} = \sum_i(\boldsymbol{w_i * Ib_i}) \quad \textbf{(5)}$$

There are 3 classic ways to calculate profiles proximity:

1) Absolute weighted differences under Norm 1 (L1 Norm): This method calculates the sum of absolute differences between the components of Profiles A and B, sensitively to each's difference's magnitude.

$$\boldsymbol{A - B} = \sum_i|\boldsymbol{a_i - b_i}| \quad \textbf{(6)}$$

2) Absolute weighted differences under Norm 2 (L2 Norm): This method, also knows as Euclidean distance, calculates the square root sum of square absolute differences between the components of Profiles A and B. This norm highlights the absolute difference squaring values, making it sensitive to each difference's magnitude.

$$\boldsymbol{A - B} = \sqrt{\sum_i(\boldsymbol{a_i - b_i})^2} \quad \textbf{(6)}$$

3) Normalized dot product (cosine similarity): This method, computes the cosine of the angle between the two vectors (A and B) to assess their similarity. This measure indicates how aligned the two vectors are. A value close to 1 implies high similarity (small angle), while a value close to 0 implies low similarity (orthogonal vectors).

$$\boldsymbol{A \cdot B} = \mathbf{a_i} * \mathbf{b_i} \,/(\sqrt{\sum_i(\mathbf{a_i})^2} \times \sqrt{\sum_i(\mathbf{b_i})^2}\,) \quad \textbf{(7)}$$

This paper aims to compare the classical pattern recognition used in AI (Dot Product and Weighted Absolute Differences) with G index, introduced by Claudio Garuti, which is defined as:

$$\boldsymbol{G(A,B)_i = W_i \times \left(\frac{min(Ia{:}i, Ib_i)}{max(Ia_i, Ib_i)}\right)} \quad \textbf{(7)}$$

Due to its nature of normalizing compatibility point to point on a scale through the minimum and maximum, G index is sensitive and accurate to weight variations.

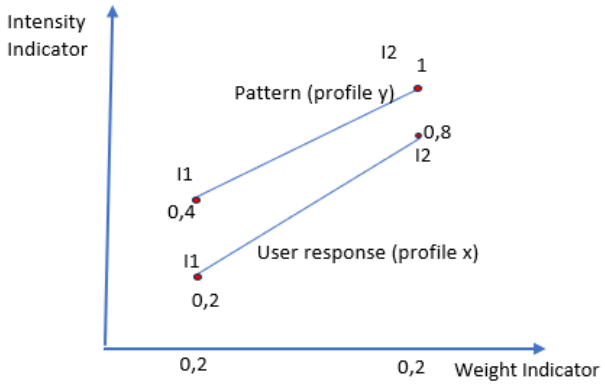
G index can be complemented by the normalized weighted distance formula, as:

$$\boldsymbol{D(A,B) = W_i \times (1 - min(Ia_i, Ib_i)/max(Ia_i, Ib_i)) = W_i - G(A,B)_i} \quad \textbf{(8)}$$

5. Results/Model Analysis

In the following, the three previously described methods will be compared in both weighted and independent scenarios by applied examples. In the case of independence from weight (both symptoms are equally important, $w_1 : w_2 = 1:1$), the scenario is highly simplified, allowing an examination of how intensity values affect variation independently of the importance of weights.

Let´s compare two profiles (Y and X) with different intensity values, in which profile Y represents the pattern to be recognized and profile X the user's response. Let $w_1 = w_2 = 0.2$ (in this case, same weight means independency from the weight), $Iy_1 = 0.4$, $Iy_2 = 0.8$, $Ix_1 = 0.2$ and $Ix_2 = 1.0$ as shown in the graph:



Then, replacing the absolute values for both intensities in equation **(6)** and **(7)**:

$$I_1 = 0.2 * (0.4 - 0.2) = 0.04$$
$$I_2 = 0.2 * (1.0 - 0.8) = 0.04$$
$$\boldsymbol{I_1 = \sqrt{0.2^2(0.4 - 0.2)^2} = 0.04}$$
$$\boldsymbol{I_2 = \sqrt{0.2^2(1.0 - 0.8)^2} = 0.04}$$

Figure 2: Comparing two weight independent profiles

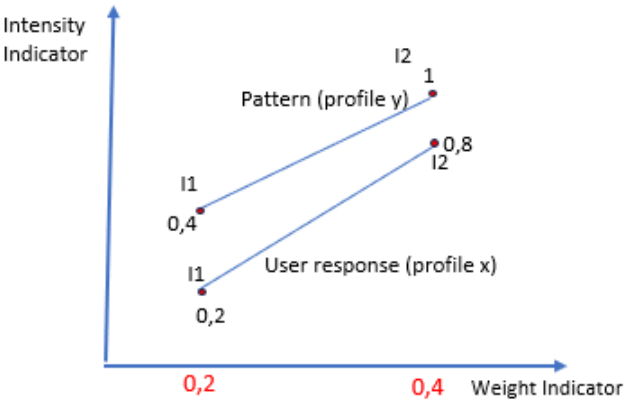
It's clear that the variation is the same for both intensities, and that the order of the intensities is irrelevant since both yield to the same variations.

However, using relative values – equation **(3)** – the variation is different for each intensity. Thus, the order of them is now relevant, as it is shown below:

$$I_1 = 0.2 * (1.0 - 0.2/0.4) = 0.10$$
$$I_2 = 0.2 * (1.0 - 0.8/1.0) = 0.04$$

For evaluating the proximity of both profiles, both G index and DP methods can be used. In this case, from the perspective of G, Profiles Y and X are not compatible (indicating low similarity), since it´s G value is 0.796 (79.6%). Whereas from DP point of view, Y and X are high compatible vectors (99,1%).

In the case of weight dependency, the importance of each symptom is different. Let's suppose a variation in weights in a ratio of $w_1 : w_2 = 1 : 2$, i.e., the second indicator is twice important than the first.



Then, replacing absolute values for intensities 1 and 2:

$$I_1 = 0.2 * (0.4 - 0.2) = 0.04$$
$$I_2 = 0.4 * (1.0 - 0.8) = 0.08$$

$$\boldsymbol{I_1 = \sqrt{0.2^2 (0.4 - 0.2)^2} = 0.04}$$
$$\boldsymbol{I_2 = \sqrt{0.4^2 (1.0 - 0.8)^2} = 0.08}$$
$$\boldsymbol{I_2 > I_1}$$

Figure 3: Comparing two weight dependent profiles

Based on absolute values, $\boldsymbol{I_2}$ has a greater variation – indicating that $\boldsymbol{I_2}$ has a larger impact in this weighted scenario – meaning that $\boldsymbol{I_2}$ should be chosen first. However,

using relative values for intensities 1 and 2 changes the result, which is shown below by replacing intensity and weight values into equation **(3)**:

$$I_1 = 0.2 * (1.0 - 0.2/0.4) = 0.10$$
$$I_2 = 0.4 * (1.0 - 0.8/1.0) = 0.08$$
$$I_1 > I_2$$

In this case, I_1 has the greater variation, meaning it should be chosen first. This way it is proved that absolute and relative values may result in completely contrary results given weighted and not weighted environments.

To understand better this situation, it's important to consider the ratios and topological similarity within the example above. Firstly, 0.4 is twice as large as 0.2, while 1 is far from being twice as large as 0.8, even though the difference is 0.2 in both cases. This illustrates that the same absolute difference can mean different things depending on the scale. Secondly, and topologically speaking, 1.0 is closer to 0.8 than 0.4 is to 0.2, even though both pairs have the same interval, emphasizing that perception of similarity can vary based on context, not just raw numerical differences. For example, in the case of dependency with weight, even though the weight of indicator I_2 is twice I_1 (i.e. twice as important),

indicator I_1 is the one that must be chosen ($0.1 > 0.08$). Unlike the traditional formulae WAD, which choose I_2 ($0.08 > 0.04$). This error happens because the weight increases the initial error produced by the absolute difference $(y - x)$, making indicator I_2 and not I_1 the chosen one.

Concerning proximity between the profiles Y and X, the G index indicates a value of 0.802 (80.2%), which is closer than the case before but still not compatible. While the DP method value is 0.997 (99.7%), indicating an almost identical pattern between both profiles, leading to a very high (and wrong) degree of similarity.

It's important to notice that for both compatibility methods the values increase from weight independent to weight dependent scenarios. Specifically, G increases in $0.006/0.796 = 0.0075$ (0.75%) and DP in $0.006/0.991 = 0.0061$ (0.61%), meaning that the ratio of change of G is slightly greater than DP. Thus, the weight's differences are better captured with G index.
Another comparison case between G and DP, numerically and graphically presented is shown in the next figure:

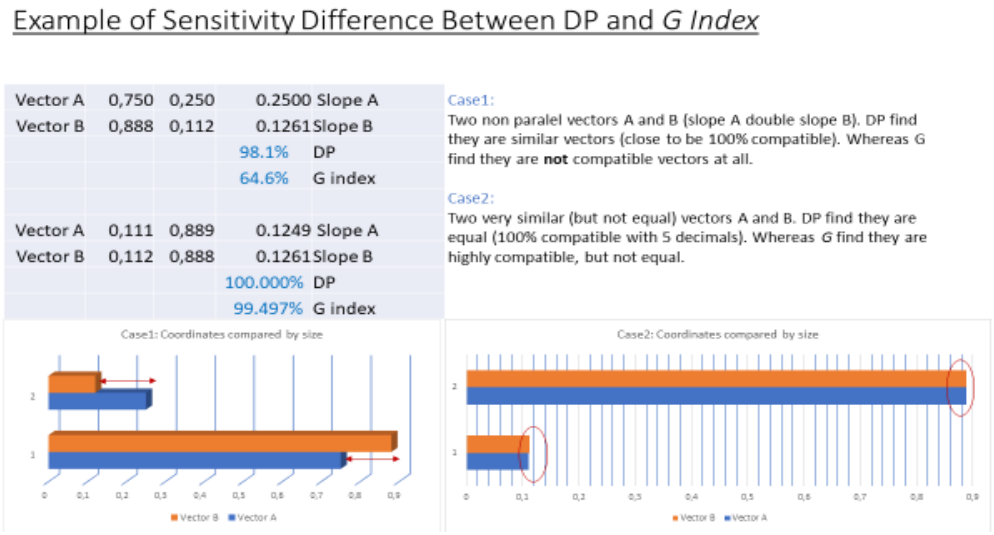


Figure 4: graphical examples of differences between G and DP

Case 1 presents two non-parallel vectors A and B (slope of A almost double slope of B). DP find they are similar vectors (close to be 100% compatible), whereas G find they are **not** compatible vectors at all. The difference between its coordinates is shown graphically in the left bar graph by the red arrow.
Case2: presents two very similar but not equal vectors, (B slope is 1,37% bigger than A slope). DP find they are equal (100% compatible with 5 decimals), whereas *G* find they are highly compatible, but not equal (A and B are not the same vector). The slightly difference between its coordinates is shown graphically in the right bar graph by the red circles.

Those differences between G and DP may come from the normalization form of each formula. While G normalization is done point to point (coordinate to coordinate) and then

sum, DP normalization is done in just one time over all the sum. Considering that each point may have different weight this different way of normalization become relevant.

Actually, G is being used in disease pattern recognition for medical diagnosis through Medical Sapiens System (www.medicalsapiens.com) a medical diagnose support system (MDSS), as shown in the next figure.

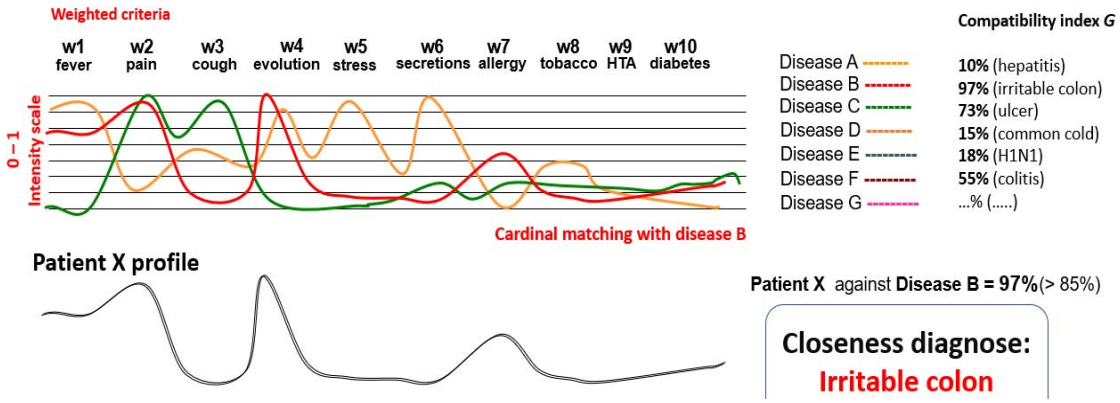


Figure 5: Application of pattern recognition in medical example

6. Conclusions

Based on the previous results, several conclusions can be drawn. First, the importance of proper measurement. In order topology (relationship between intensity values), it's important to account for intensities on a ratio scale, where differences are interpreted in relation to the values themselves, not just as absolute values. This way, a poorly executed measurement procedure can lead to errors and slowdowns in calculations, especially when the weight of each indicator is factored in. This is why WAD is not an adequate method for calculating distance in weighted environments, as it fails to capture the nuance of ratio-based relationships and can lead to incorrect conclusions about proximity.

Secondly, normalized Dot Product performs better than Weighted Absolute Distance, but it still falls short compared to G index, since the latter is more sensitive and accurate in representing compatibility, especially in profiles with varying weights mainly due to the normalization procedure. Additionally, G is also easier and faster to calculate compared to DP, making it a practical choice for large datasets with complex weight structures.

By evaluating D as the difference between weights (W) and G for each indicator, a better calculation of the deviation between the pattern and user profiles is achieved. If this approach is applied to all indicators, the following general formula can be used:

$$\boldsymbol{Dmax = Max(W - G)_i \quad for\ i = 1, n\ indicators}$$

This formula, called the Maximum Difference Formula, identifies the point with the maximum deviation between the pattern and user profiles, and it helps pinpoint where the largest difference between the pattern and user response occurs (considering the pair

weight and intensity in each point of the profile), offering an efficient feedback mechanism for improving proximity calculations. Providing, this way, a straightforward and powerful tool for optimizing a database of disease profiles, making it easier to improve the accuracy and relevance of stored information.

7. Limitations

This study's findings may be limited by the specific datasets and examples used, which could affect generalizability. The reliance on predefined weights introduces subjectivity, as weight determination varies by context. Future research should include broader use cases, real-world testing, and exploration of adaptive weight assignment to enhance the model's reliability and applicability.

8. Key References

Chen, L., & Du, Y. (2008). *Analysis of weighted measures in clinical pattern recognition.* Computational Medicine and Health.

Garuti, C. (2014*). Measuring in Weighted Environments*. International Journal of AHP (IJAHP) 2014.

Garuti, C. (2012*). Measuring in Weighted Environments: Moving from Metric to Order Topology*. Santiago, Chile: Universidad Federico Santa Maria (Book, 72pp).

Garuti, C. (2014). *Compatibility of AHP/ANP vectors with known results. Presentation of a suggested new index of compatibility in weighted environments.* International Symposium of the Analytic Hierarchy Process.

Garuti, C. (2007). *Measuring Compatibility in Weighted Environments: When close really means close?* International Symposium on AHP, 9, Viña del Mar, Chile. 2007.

Garuti C. 2021, *"How to Obtain a Global Reference Threshold in AHP/ANP"*. IJAHP International Journal of AHP Vol. 13. N°1. Abstract 533 | PDF Downloads 166 | DOI https://doi.org/10.13033/ijahp.v13i1.802

Garuti C. 2022, *"A Set Theory Justification of Garuti´s Compatibility Index: generalization of Jaccard index working within a weighted environment*". Advances in Mathematics Research. Vol.30 e-ISBN: 978-1-68507-536-1. DOI: https://doi.org/10.52305/BDFQ3979. Editor: Albert R. Baswell. Published by Nova Science Publishers NY.

Kononenko, I. (2001). *Machine learning for medical diagnosis: history, state of the art, and perspective*. Artificial Intelligence in Medicine.

Saaty, T. L. (2005). *Decision making with the Analytic Hierarchy Process*. International Journal of Services Sciences.

Singh, K. P., Gupta, M. (2010). *Cosine similarity and its applications in clustering for high-dimensional data*. Journal of Computer Science.