Original Scientific Paper

Journal of Applied Engineering Science

doi:10.5937/jaes17-21960

Paper number: 17(2019)4, 632, 468 - 472

BASIC CONCEPT PYTHAGORAS TREE FOR CONSTRUCT DATA VISUALIZATION ON DECISION TREE LEARNING

Erlin Windia Ambarsari[1], Aulia Ar Rakhman Awaludin[1], Andri Suryana[2], Purni Munah Hartuti[1], Robbi Rahim[3]*

[1]Engineering and Computer Science Faculty, Universitas Indraprasta PGRI, Indonesia

[2]Post Graduate of Universitas Indraprasta PGRI, Indonesia

[3]Sekolah Tinggi Ilmu Manajemen Sukma, Indonesia

Decision Tree in Data Mining frequently used to learn the pattern by interpreting data. A hierarchy of tree model in Decision Tree as data visualization which often used makes fully load space. Another option in using model is Phytagoras Tree. Pythagoras Tree in this study is the basic concept of Pythagorean Theorem that used by a binary hierarchy with a fractal technique which the shape using the square as branches enclose a right triangle. A fractal of Pythagoras Tree is the dataset which split the subsets into trunks and leaves. Construct a fractal of Pythagoras Tree depends on the angle θ for build branches followed by square area. Pythagoras Tree model is an easy way to understanding the dataset based on the size of the square. The smaller the size, the fewer instances in the rectangle. Also, data associations easily traced when filled with color.

Key words: trees, fractals, decision trees, construction

INTRODUCTION

Data Mining is useful in learning pattern big data to understand the flow of the problem for the base of further decision. The techniques of the pattern as a data mining task for analyzing are descriptive, predictive, and prescriptive. One of the models which often used for predictive in Data Mining is Decision Tree. It serves by predicting attributes used as branches based on the target as part of the leaves. Decision Tree used for covering classification and regression with visualization for interpretation data in decision analysis. The Decision Tree Learning represented a tree model which makes the rule of IF-THEN or Association. A tree model has a role as a visualization that describes how the dataset is branched. Therefore, branches making it easier to trace interconnected data until dataset become leaves which it can not split again.

Based on the study of [1], a tree model with hierarchies which often used makes run out of space to visualization. The model becomes full and complex when branches are going to depth. Therefore, the alternative option that used is Pythagoras Tree, which based on Pythagorean Theorem. Pythagoras Tree of the first introduction by [2] that a tree model used a binary hierarchy with a fractal technique which the shape using the square as branches enclose a right triangle. A fractal of Pythagoras Tree is the dataset which split the subsets into trunks and leaves. A fractal becomes recursive until all dataset has been separate.

Therefore, we observe the Pythagoras Tree in Decision Tree Learning based on Pythagorean Theorem. Pythagoras Theorem had over 371 proofs by discoverers [3] which several visualizations can use to build the tree based on the dataset, such as in the study of [4] using the central square theory [5] form branches. Branches itself is using Geometrical Pythagoras' and Plato's visualizations by construct data.

Whereas the concept built tree in this study is to construct data visualization based on Pythagoras' Geometric with using a right-angled triangle [6], and Algebra [7] which square made with four right triangles. Both Pythagoras Theorem Proofs applied on Decision Tree Training. The Geometric is used to separate dataset and determine an angle in a right-angled triangle. The function is to make a tree model. Also, the algebraic used for measuring the number of datasets in a square. Therefore, the data visualization depended from the square's size and an angle in a right-angled triangle, until the result is a fractal of Pythagoras Tree.

Determined Phytagoras Tree uses the Decision Tree Algorithm, which has several methods that used as an experiment: ID3 [8], CART [9], and C4.5 [10]. In this study, we used ID3 with Standard Deviation Reduction (SDR) for discovering regression which the dataset is a numerical type. SDR used as branches separation determination for construct Pythagorean Tree as data visualization.

METHODOLOGY

Pythagorean theorem

Similar to the name, Pythagoras who discovered the right triangle which has two legs and one hypotenuse. Two legs in right triangle are Opposite and Adjacent. The opposite is across to the angle θ, Adjacent is next to the angle θ, and the hypotenuse has the longest distance, which is the sum of the opposite and adjacent. Also, the

*usurobbi85@zoho.com

value of hypotenuse always the same when the angle θ changes, because of the side of the opposite has a 90-degree angle. The equation of edge for a right triangle is:

$$\sin\theta = \frac{Opposite}{Hypotenuse}, \cos\theta = \frac{Adjacent}{Hypotenuse} \quad (1)$$
$$\tan\theta = \frac{Apposite}{Adjacent}$$

Pythagoras Theorem represented in terms of area, which the hypotenuse square area is equal to the sum of the opposite and adjacent square area. It illustrated as Figure 1 as below:

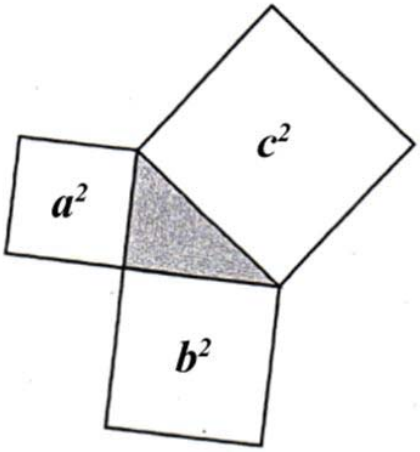

Figure 1: Square area of the right triangle [11]

The relationship between hypotenuse (c), opposite (a), and adjacent (b) as the equation is:

$$c^2 = a^2 + b^2 \quad (2)$$

Equation 2 was proof with the square area in Figure 1, was surrounded by four right triangles, illustrated in Figure 2, which is similar to comparison 5. Therefore, Decision Tree obtained from dataset c^2 and subset a^2 and b^2, which is split based on SDR.

$$Area1 = (a+b)^2 \quad (3)$$

$$Area2 = h^2 + 4\left(\frac{1}{2}ab\right) \quad (4)$$

$$Area1 = Area2$$
$$(a+b)^2 = h^2 + 4\left(\frac{1}{2}ab\right)$$
$$a^2 + 2ab + b^2 = h^2 + 2ab \quad (5)$$
$$a^2 + b^2 = h^2$$

Standard Deviation Reduction ID3

SDR based on Standard Deviation (σ) which to calculate a numerical dataset until itcontains instances the homogeneity values completely. The σ equation is [12]:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \mu)^2} \quad (6)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} u_i \quad (7)$$

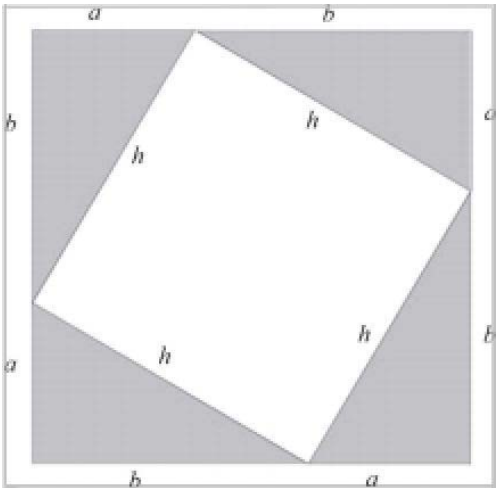

Figure 2: Square constructed with four right triangles [7]

The standard deviation isbeneficial to quantifythe spread dataset while the mean (μ) used asa dataset concentration measurement, which able to describe a data in the range of the dataset. The mean is not able to measure of concentration for nominal and ordinal data types.

Therefore, the steps that can do with SDR are as follows:

- Calculate the standard deviation for the target's attributesas the basis of the decision.
- Separate the datasetto examine the relationship two attributes variable: target and predictor with equation σ(T,P). The predictor is able used to determinethe hypotenuse and two legs (opposite and adjacent) values.
- Compute the standard deviation for each branch. The result of the standard deviation, which subtracted from the standard deviation before separate to gain SDR.

$$SDR = \sigma - \sigma(T,P) \quad (8)$$

where

$$\sigma(T,P) = \left(\frac{n_1}{n}\, x\, \sigma(P)_1\right) + \left(\frac{n_2}{n}\, x\, \sigma(P)_2\right) \quad (9)$$

The attribute with the most significant SDR values is the best choice.

Construct decision tree recursively until the Coefficient of Variant (CV) become smaller than the threshold which is determined.

$$CV = \frac{\sigma(P)}{\mu} x\ 100\% \quad (10)$$

RESULT AND DISCUSSION

Data visualization of Decision Tree done made by the separation of datasets using SDR. Constructing the Pythagoras Tree is done by obtaining the number of instances that used as hypotenuse and legs. Table 2 is an example of a data sample where the total number of datasets for instances is n = 114 with SDR selected as the best choice, which the name attributes initialused an alphabet. As an example for SDR understanding can use some raw data in Tabel 1 as below.

Table 1: The raw sample data within stances n = 114[13]

Predictors					
n	D	A	B	C	Target
1	0.985673	897.67948	20	1530	31500
2	0.952585	620.46443	22	245	28350
3	0.948748	617.97662	21	836	25200
4	0.964798	605.09624	20	501	28350
5	0.925402	596.66193	20	1017	21000
6	0.963987	580.92263	22	460	18900
7	0.776459	496.49018	21	2101	0
8	0.980644	454.78552	21	470	18900
…	…	…	…	…	…
106	0.942809	163.10596	0	33	0
107	0.978727	158.71686	5	66	0
108	1	158	1	4	0
109	0.958445	157.45874	4	47	0
110	0.533533	157.34947	20	714	0
111	0.973679	156.35663	6	203	0
112	0.992839	153.9314	11	161	0
113	0.882369	144.2674	0	37	0
114	0.940258	137.27765	3	26	0

The first step to a determined decision tree is to find μ, which is the overall total attribute target/114 instances, based on equation 7.It is means (31500+28350+25200+28350+… +n114)/114 = 6138.816. Calculation σ by subtracting each attribute target with μ and result of the value use power of two, which then added together. Also, it divided 114 instances and make a square root, which means σ=((31500-6138.816)2+(28350-6138.816)2+…+(n114+ μ)2)/114= √48375220 = 6955.23.

Split the dataset to examine the relationship attributes of target and predictor. For assumption, we splitthe A attributes predictor n = 1 to n = 6 and n = 7 to n = 114, which means it is split instances n ≥ 496.49018 and n<496.49018. Therefore, the total of datasets for attribute target n ≥ 496.49018 is 6 instances (31500, 28350, 25800, 28350, 21000, 18900) and n ≥ 496.49018 is 108 instances.

Decision Tree does some training by comparing SDR whichever is greater. Also, when the CV is higher than 10%, the dataset needs to be subdivided; otherwise the dataset becomes leaves. In this case, instances are 108 and 6, which use for legs in a right triangle is shown in Table 2.

Table 2: Hypotenuse (n) and two legs based on SDR by three depth

Spit by Attributes	n	μ	Standard Deviation	Leg1	Leg2	Depth (Recursive)
A	114	6138.816	6955.23	108	6	1
A	6	25550	4399.432	4	2	2
A	4	28350	2227.386	0	0	2
A	2	19950	1050	0	0	2
B	108	5060.417	5281.241	38	70	2
C	38	248.684	1278.838	2	36	3
D	70	7672.5	4770.029	15	55	3

Based on table 2 that the value obtained on SDR depends on the number of instances in the dataset.Therefore, the depth to one is the overall dataset n = 114, and legs have 108 and 6. Dataset n is hypotenuse which the sum of two legs: 108 + 6. If a dataset is a square area under the Pythagoras Theorem, then c^2 = 114 followed by a^2 = 108 and b^2 = 6.

Construct a fractal of Pythagoras Tree also depends on the angle θ for build branches followed by square area. Therefore, it is essential to determine which one of the two legs are opposite or adjacent. For example, to construct a tree trunk, adjacent is more used because this part is closer to hypotenuse when compared to the opposite.

According to equation 1, the value used is not a square area, but the value of each side, which based on Figure 2, where each square area is the length of a right triangle. Therefore, hypotenuse c = √114 = 10.6770782520313. While the adjacent is gain from the other leg, 108>6 which means adjacent a = √108 = 10.3923048454133 and opposite b = √6 = 2.4494897427832.

The calculation the angle θ used from one of sin θ, cos θ, or tan θ, the value is the same because it used a right triangle which the upright side has a 90-degree angle. Therefore, the value the angle is sin θ = 2.449489742 7832/10.6770782520313 or θ = sin-1 (2.4494897427832 /10.6770782520313). The result is 13.2626760083048°.

According to Figure 3, < BAC = 13.2626760083048o and <ACB = 90o. Therefore, to obtain <ABC, reduce the value of <ACB to <BAC, 90o-13.2626760083048o = 76.73732399169520. According to Figure 1, c^2 is an area of square 114, where the circumference of the rectangle is √114, which illustrated in Figure 4. The square area also applied to legs of a^2 and b^2.

Perform recalculation instances for values of other legs. For example,the instances of 108 and 6 until there are no more legs. 6 instances divide to 4 and 2: where square area c^2 = 6; hypotenuse c = 2.4494897427832; adjacent a = 2; opposite b = 1.4142135623731; and θ = 35.2643896827547o.A fractal Pythagoras Tree is

Journal of Applied Engineering Science Vol. 17, No. 4, 2019
ISSN 1451-4117

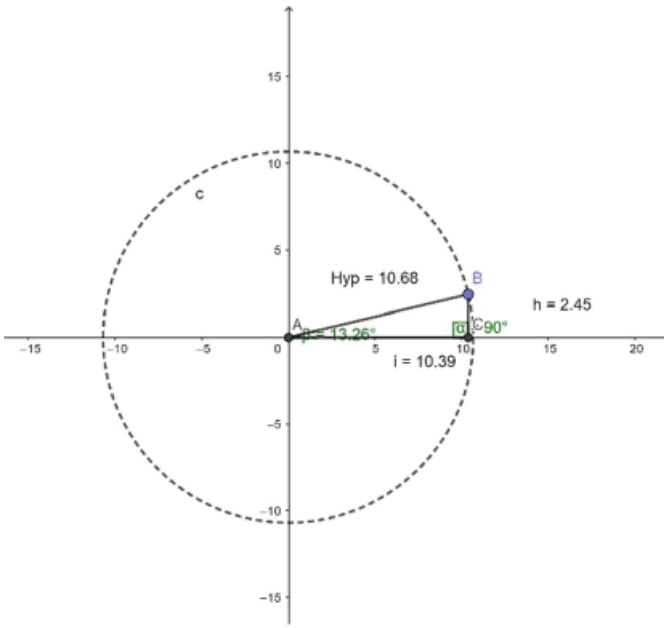

Figure 3: Construct the right triangle

recursive until the values become the homogenous instances (Figure 6).

In fundamental, datasets (assumption as black spot) for dataset itself had a classification in the square area such as Figure 5. It means the size of the square area is equal to the amount of data. It is call construct data visualization.

Illustrated by Pythagoras Tree is an easy way to understanding the dataset based on the size of the square. The smaller the size, the fewer instances in the rectangle. Also, data associations easily traced when filled with color, as in Figure 6,which shows the relationship data of IF-THEN based on the μ values.

Based on [14] concept, it is establishing a gradient color scale from 0023BF hex color (μ = 0) to D1E100 hex

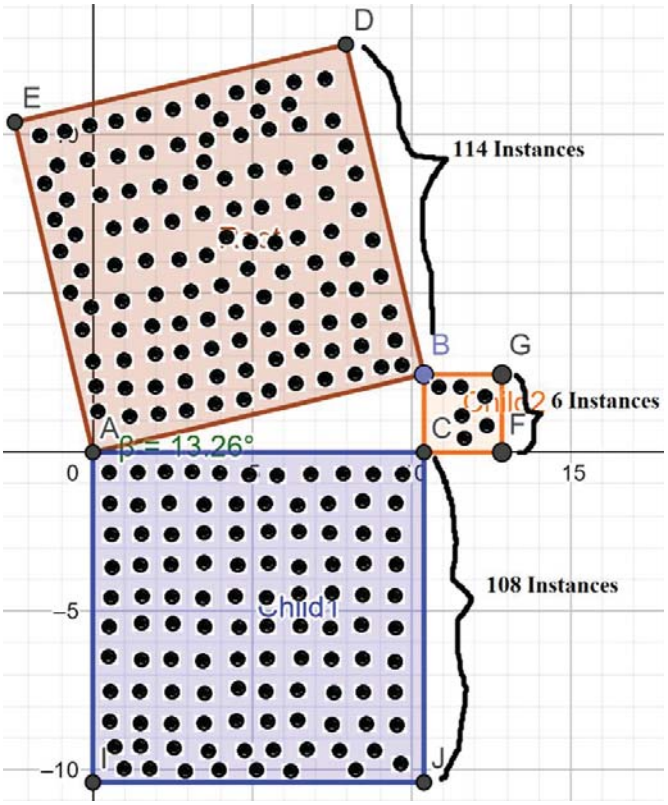

Figure 5: The dataset in Square Area

color (μ = 31500). It used to present μ values, as shown in Figure 6. For example, if μ = 6138.816 which color is 1436AB hex color. Therefore, Pythagoras Tree presents a view that simple and understandable to read.

Based on [14] concept, it is establishing a gradient color scale from 0023BF hex color (μ = 0) to D1E100 hex color (μ = 31500). It used to present μ values, as shown in Figure 6. For example, if μ = 6138.816 which color is 1436AB hex color. Therefore, Pythagoras Tree presents a view that simple and understandable to read.

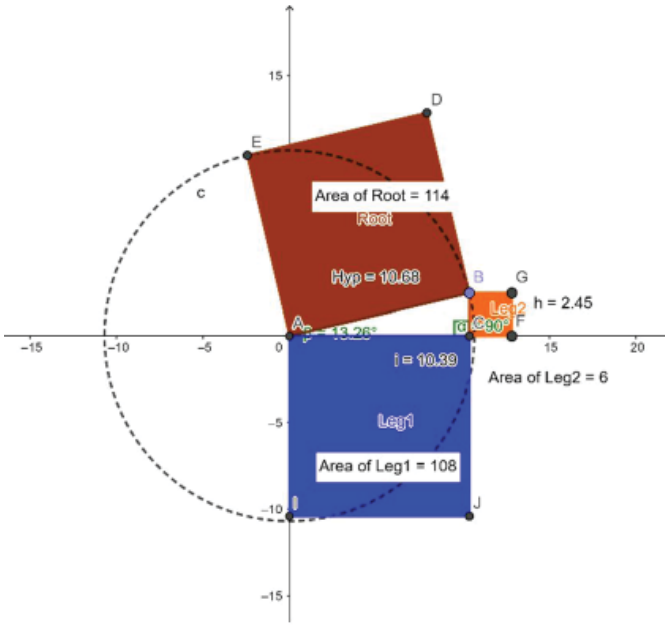

Figure 4: Pythagoras Tree on Depth to one

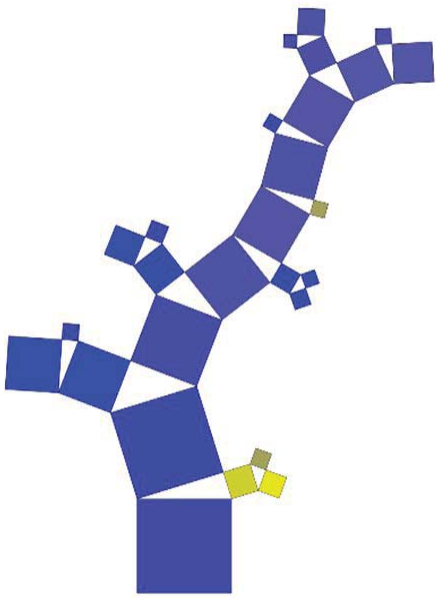

Figure 6: Pythagorean Tree with the μ color

CONCLUSION

Decision Tree with hierarchies model frequently fully loaded space in data visualization. Especially the branches are going to depth. Therefore, the alternative option that used is Pythagoras Tree, which based on Pythagorean Theorem.

Pythagoras Tree in this study used a binary hierarchy with a fractal technique which the shape using the square as branches enclose a right triangle. A fractal of Pythagoras Tree is the dataset which split the subsets into trunks and leaves. A fractal becomes recursive until all dataset has been separate. Construct a fractal of Pythagoras Tree depends on the angle θ for build branches followed by square area.

Pythagoras Tree model is an easy way to understanding the dataset based on the size of the square. The smaller the size, the fewer instances in the rectangle. Also, data associations easily traced when filled with color.

REFERENCES

1. F. Beck, M. Burch, T. Munz, L. Di Silvestro, and D. Weiskopf, “Generalized Pythagoras Trees for Visualizing Hierarchies,” in Proceedings of the 5th International Conference on Information Visualization Theory and Applications, 2014, pp. 17–28.
2. A. E. Bosman, Het wondere onderzoekingsveld der vlakke meetkunde. Breda: N.V. Uitgeversmaatschappij Parcival, 1957.
3. B. Ratner, “Pythagoras: Everyone knows his famous theorem, but not who discovered it 1000 years before him,” J. Targeting, Meas. Anal. Mark., vol. 17, no. 3, pp. 229–242, Sep. 2009.
4. L. Teia, “Anatomy of the Pythagoras’ tree,” Aust. Sr. Math. J., vol. 30, no. 2, pp. 38–47, 2016.
5. L. T. Gomes, “Pythagoras Triples Explained via Central Squares,” Aust. Sr. Math. J., vol. 29, no. 1, pp. 7–15, 2015.
6. V. Dlab and K. S. Williams, “The Many Sides of the Pythagorean Theorem,” Coll. Math. J., vol. 50, no. 3, pp. 162–172, 2019.
7. J. R. Parada-Daza, M. I. Parada-Contzen, J. R. Parada-Daza, and M. I. Parada-Contzen, “Pythagoras and the Creation of Knowledge,” Open J. Philos., vol. 04, no. 01, pp. 68–74, Jan. 2014.
8. J. R. Quinlan, “Induction of Decision Trees,” Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.
9. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification And Regression Trees. Routledge, 2017.
10. J. R. Quinlan, C4.5: Programs for Machine Learning. 1993.
11. S. Swaminathan, “The Pythagorean Theorem,” J. Biodiversity, Bioprospecting Dev., vol. 01, no. 03, pp. 1–4, Sep. 2014.
12. M. F. Al-Saleh and A. E. Yousif, “Properties of the Standard Deviation that are Rarely Mentioned in Classrooms,” AUSTRIAN J. Stat., vol. 38, pp. 193–202, 2009.
13. E. W. Ambarsari, S. Khotijah, and L. Sunarmintyastuti, “Pemodelan Reward Rule Game Streamer Indonesia Tingkat Amatir dengan Orange Data Mining,” STRING (Satuan Tulisan Ris. dan Inov. Teknol., vol. 4, no. 1, pp. 9–17, Aug. 2019.
14. W. A. C. Rojas and C. M. Villegas, “Graphical representation and exploratory visualization for decision trees in the KDD process,” Proc. - 2012 9th Electron. Robot. Automot. Mech. Conf. CERMA 2012, vol. 73, no. Dm, pp. 203–210, 2012.

Paper submitted: 01.06.2019.
Paper accepted: 06.10.2019.