THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

EXTRACTING A POWERFUL INDICATOR BASED ON DAN BARBILIAN APOLLONIAN METRIC USING COLON TISSUES GENE EXPRESSION MATRIX IN ORDER TO IMPROVE PERFORMANCE IN PREDICTION AND CLASSIFICATION OF COLON TUMOUR

Alexandru DAIA[1] and Constantin IONESCU-TÎRGOVIŞTE[2]

[1] Upwor INC
[2] National Institute of Diabetes, Nutrition and Metabolic Diseases, NC Paulescu, Bucharest, Romania
Corresponding author: Alexandru DAIA: alexandru130586@yandex.com;

Accepted 29 May 2019

Having a data set with gene expressions from colon cancer, we aimed to construct a new feature to improve machine learning classification task.

Key words: feature extraction, gene expression, machine learning, logistic regression, Dan Barbilian, data science, colon cancer, genes, associated genes.

INTRODUCTION

Generally biology and human pathology, in particular, operate with tissues or organs (several dozen) with organ structures (millions), cells (billions) and molecules (trillions). Among different members of some categories (diabetic and non-diabetic – for example) there are large differences, which explains the great heterogeneity of clinical forms. Within these clinical forms (paediatric diabetes, adult diabetes, senile diabetes) adds another type of heterogeneity that ultimately makes it impossible to identify two patients, even in the monozygotic twins. In these conditions the prediction of diabetes became a stringent necessity due to the current epidemic of diabetes that affects half a billion people. However the prediction of diabetes is difficult due to the high number of organs and tissues which take part in the homeostasis of the human body.

To exemplify how is regulated this important function of the body we will mention only the regulation of the blood glucose in the normal individual[1]. For this, it is important to understand the endocrine function of the pancreas and the secretion of insulin, a hormone discovered by N.C. Paulescu in 1921[2]. So we have to make a brief presentation of the pancreas structure. The pancreas is a relatively large organ (70 g in adults), which comprises two substructures, completely different and functioning independently. Exocrine tissue (which secretes enzymes required for digestion and absorption of nutrients in intestines) is the pancreatic exocrine acines (about 97% of the whole). The remaining 1.5–3%, ensuring the internal secretion of the pancreas, is represented by small sub-structures called the Langerhans Islands, following the name of the researcher who discovered them in 1869[3]. The 2–3 grams of tissue representing these islets (about 2 million) are distributed randomly across the pancreas. In each of these islands there are approx. 5000 cells including β-pancreatic insulin secreting cells (about 68–70%, about 3000 / island). Making a simple calculation would result in 6×10^9 of beta-pancreatic cells. In each of these cells there are approx. 12.000 secretory vesicles, and in each of them is stored approx. 200.000 insulin molecules. It results in a number of difficult to imagine molecules (14.4×10^{17}). It should be noted that these molecules are similar but not absolutely identical. Insulin molecules act on all insulin-dependent cells in the body, including adipocytes, hepatocytes, and muscles whose number is difficult to perceive[4,5].

To process such "big data", inclusively in order to predict diabetes, the only reasonable solution is to use information technology in an appropriate manner. This is also, the purpose of this work, which doesn't refer to metabolic pathology but to

another important topic: the oncological pathology. For that we will refer to the genes expression in colon cancer, having the advantage of the access to an Open Source database.

Our main objective of this research is to obtain a superior method of extracting a new feature meaning useful information from existing predictors in a colon cancer study.

Feature extraction is unbounded and depends very much on data particularities. On previous experiments and research we have used Octav Onicescu information energy called also kinetic energy for various data science task including feature extraction as could bse seen in references [8] and [9].

Creating new features (predictor) is part of a process called feature extraction. Feature extraction is a process with unbounded possibilities according to data set particularities that aims to use existing data set in order extract new features that are influential for the prediction task in order to improve performances of the machine learning models.

In the past in our previous research we have also introduced a method of feature extraction using a method of another great Romanian mathematician and staticistician, expert in probabilities called *information energy* or also called *kinetic energy*.

For kinetic energy we used it to as method of feature extraction on natural language processing in order to improve accuracy of classifier for text data of different authors by text content as features and author owner as the target.

DATA AND METHOD

In 1999, Uri Alon[6,7] a highly cited researcher on that focuses among others on expression of genes currently working at Weizmann Institute of Science from Israel analyzed gene expression data for 2,000 genes from 40 colon tumour tissues and compared them with data from colon tissues belonging to 21 healthy individuals, all measured at a single time point. We can represent his data as a 2,000×61 gene expression matrix, where the first 40 columns describe tumour samples and the last 21 columns describe normal samples.

Having 2000 features and 81 cases. We decided to try a Logistic Regression in order to see initial performance meaning balanced accuracy and diagonal of the confusion matrix[8–13].

For a deeper understanding of the starting point of present paper, some biographical information about mathematician Dan Barbilian are necessary. Dan Barbilian was prominent mathematician born on 18 march 1895 died on 11 august 1961 with very good achievements in mathematics such as contributions in Ring Geometry[14], Barbilian Planes[15] and theories of modern algebra, post-mortem member of the Romanian Academy[16].

Published in 1934, the research of mathematician Dan Barbilian describing Apollonian metric will be quoted in several scientific articles and aims to calculate the Euclidean distance between point a and point b starting from the metric of the k region Figure 1.

Figure 1

$$d(a,b) = \log \max_{p \in J} pa / pb) + \log \max_{q \in J} qb / qa)$$

His theory about Barbilian Spaces was presented into four chapters and over time there where researched by a vast number of people from academia world.

EXPERIMENTING NUMBER 1

The initial experiment was using a baseline logistic regression to perform the binary classification where the outcomes are in the set {0,1} where:

- 0 means healthy patients
- 1 patient with colon tumour

The reason why we chooses Logistic Regression instead of other classification method is because the data set is small and other models such as deep neural networks or decision tree based models (ensembles such as Random Forest, Gradient Boosting Trees) are to non-linear and on small data non linear tend to encounter much over fitting.

The results of logistic regression on data without the feature engineering we propose was the following:

Balanced accuracy score (macro-average of recall scores per class): **0.92**.

In this experiment the chances to predict that a particular set of gene expressions of a person to really have colon cancer are **96%** meaning **3%** are predicted wrong to not have colon cancer while in reality they have.

On the other hand the chances that a particular set of gene expressions of a person to be predicted not to have colon cancer are arround 92% meaning 8% are predicted wrong to have colon cancer while in reality they do not have.

EXPERIMENT NUMBER 2

In experiment number we created a new feature based on Dan Barbilian apollonian metric. In order to be able to create a new feature aimed not to apply exactly the apollonian metric as it is described but to simulate as much as possible the use of the metric we did the next steps which creates other variables that will be filled in the formula of the distance:

i. As could be seen in Figure 2 we first create 3 clusters of the data using k-means clustering from sklearn, an open source machine learning library designed to work wih Python programming language. For each of the 61 cases we made predictions in order to assign corresponding cluster.
ii. For each of the clusters we computed also the centre of gravity.

We define the **centre of gravity** of Data as the point whose i-th coordinate is the average of the i-th coordinates of all points from Data. For example, the center of gravity of the points (3,8), (8,0), and (7,4) is [(3+8+7)/3,(8+0+4)/3]=(6,4).

iii. We create 2 components T-distributed Stochastic Neighbour Embedding (t-SNE) a method for dimensionality reduction of the data.

From Figure 2 where is the definition of the metric we assign the values of the variables from formula according to our data asset as follows:

a) a is current row gene expressions meaning the features that initially had;
b) b is centre of gravity of corresponding cluster;
c) p and q first and second component of Distributed Stochastic Neighbour Embedding (t-SNE), where p is first component column T1 and q is second component denoted T2 in the table.

The results of logistic regression on data with currently created feature based o Barbilian formula of distance improved from previous expression as follows:

Balanced accuracy score (macro-average of recall scores per class): **0.96**.

In this experiment the chances to predict that a particular set of gene expressions of a person to really have colon cancer are **100** meaning **30%** are predicted wrong to not have colon cancer while in reality they have.

On the other hand the chances that a particular set of gene expressions of a person to be predicted not to have colon cancer are around 92% meaning 8% are predicted wrong to have colon cancer while in reality they do not have. Similar with experiment number 1 in this case.

$$d(a,b) = \log \max_{p \in J}(pa/pb) + \log \max_{q \in J}(qb/qa)$$

	0	1	2	3	...	1996	1997	1998	1999	Cluster	Gravity	T1	T2	max_p	max_q
0	9164.253700	6719.529500	4883.448700	3718.158900	...	152.195000	186.567500	44.472500	16.773750	0.000000	320.988111	-207.106995	7.819423	11087.208678	0.127274
1	6246.448700	7823.534100	5955.835000	3975.564300	...	48.338750	42.520000	49.982500	16.085000	0.000000	320.988111	-102.999519	-130.024521	8833.221011	0.154728
2	2510.325000	1960.654500	1566.315000	3072.816100	...	5.878750	36.297500	9.815000	21.883750	0.000000	320.988111	283.088867	248.681152	5238.801343	0.235866
3	4028.710000	3156.159100	2870.255000	4417.591100	...	24.196250	29.766250	44.376250	52.290000	0.000000	320.988111	379.811707	249.085587	8248.025196	0.163908
4	5271.517500	4740.768200	3318.513700	6792.348200	...	43.230000	67.150000	65.610000	44.042500	0.000000	320.988111	-67.052979	-56.468067	11344.019981	0.124752
5	14173.054000	8411.861400	6042.840000	8766.046400	...	163.760000	205.438750	58.243750	44.057500	1.000000	626.908827	-228.015747	-80.433235	191.508849	0.104000
6	4985.218800	4735.793200	4075.122500	2845.248200	...	32.702500	33.491250	38.727500	21.631250	0.000000	320.988111	194.942642	369.552429	9048.723209	0.151602
7	5627.251200	3619.531800	2606.500000	2544.451800	...	9.187500	18.015000	8.447500	37.822500	0.000000	320.988111	317.797729	-82.400330	830.689384	0.140920
8	4865.220000	3237.777300	2341.086300	2372.680400	...	34.633750	33.043750	36.897500	64.521250	0.000000	320.988111	248.557770	174.806503	5740.349249	0.219784
9	4412.476300	1752.395500	1280.323800	5200.344600	...	26.408750	32.325000	39.608750	36.042500	0.000000	320.988111	324.073090	172.989899	7095.450236	0.185595
10	6995.410000	6658.109100	5316.396200	4047.514300	...	104.941250	127.938750	104.917500	59.231250	0.000000	320.988111	-148.274185	-45.287243	11832.297756	0.120222

Figure 2

CONCLUSIONS

Feature extraction on biological data when using machine learning methods is a step that is fundamentally necessary before training machine learning models and our method of feature extraction using Dan Barbilian apollonian metric given a proof of concept that is good for improving prediction performance.

From Barbilian theories there are other experiments that could be performed and tested on other data sets or even in other domains starting from machine learning on structured data (row data) and going even beyond to unstructured data such as natural language processing or image processing.

REFERENCES

1. Ionescu-Tirgoviste C., *Prolegomenon to the European Constitution Book of Diabetes Mellitus*. Proc. Rom. Acad., Series B, *3*, p. 179–213, 2008.
2. Paulescu N.C.: *Recherche sur le rôle du pancréas dans l'assimilation nutritive*. Liège, p. 85-109, 1921.
3. Langerhans P. *Beitrage zur mikroskopischen Anatomie der Bauchspeicheldruse*. Med. Diss, Berlin, 1869.
4. Bain R. J., Stevens D. R., Wenner R. B., Ilkayeva O, Muoio M. D., Newgard B. C.: *Metabolics applied to diabetes research moving from information to knowlege*. Diabetes 58: 2429 - 2443, 2009
5. Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S; DIAGRAM Consortium, Meigs JB, Altshuler D, Knowler WC, Florez JC; Diabetes Prevention Program Research Group: *Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program*. Diabetes. 60(4):1340-1348, 2011.
6. Alon U, N Barkai, DA Notterman, K Gish, S Ybarra, D Mack, AJ Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligo nucleotide arrays*. Proceedings of the National Academy of Sciences 96 (12), 6745-6750, 1999.
7. Alon U. *Biological networks: the tinkerer as an engineer* Science 301 (5641), 1866-1867. 2003.
8. *Experimented kinetic energy as features for natural language classification*, Proceedings of Romanian Academy,Issue 3, Volume 20, 2018.
9. https://www.researchgate.net/publication/330042218_EXPERIMENTED_KINETIC_ENERGY_AS_FEATURES_FOR_NATURAL_LANGUAGE_CLASSIFICATION?_sg=7-bvKaWBImtiElUV4mkhL6QHymbhs8xMbU0C9wVJ5MAv37ovlj3ZoblJ55F7M4UiYo33lpbYooD0PuFmoURlVR822CdyxrujY9JipCfr.FP5S_fqN-E7mkhJ4QcW_9b_vPW0p0utvAlSQh3zjpavwxltEKaOwkkcH3tNVzbfMtALi6YL735pTtcwlVyZaKQ
10. https://www.researchgate.net/publication/326069140_Onicescu_ kinetic_energy_and_applications
11. https://en.wikipedia.org/wiki/Ion_Barbu
12. https://en.wikipedia.org/wiki/Feature_extraction
13. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
14. Rings and Geometry R. Kaya, P. Plaumann, K. Strambach - 2012 - Mathematics
15. The Geometrical Barbilian's Work from a Modern Point of View, Radu Iord, 1996.
16. https://en.wikipedia.org/wiki/Ion_Barbu