EPiC Series in Computing

Volume 69, 2020, Pages 230–241

Proceedings of 35th International Conference on Computers and Their Applications

Optimize Neural Network Algorithm of Missing Value Imputation for Clustering Chocolate Product Type Following "STEAMS" Methodology

Mason Chen[1], and Charles Chen[2]
[1] Stanford University OHS, Palo Alto, USA
[2] Morrill Learning Center, San Jose, USA
Mason05@ohs.stanford.edu, Charles.chen.training@gmail.com

Abstract

A "STEAMS" (Science, Technology, Engineering, Artificial Intelligence, Math, Statistics) approach was conducted to handle the missing value imputation of clustering Chocolate Science patterns. Hierarchical clustering and dendrogram analysis were utilized to cluster the commercial chocolate products into different product groups which can indicate the nutrition compositions and product health. To further handle the missing value imputation, a neural network algorithm was utilized to predict the missed Cocoa percentage (Cocoa%), based on other available nutritional components. The Hyperbolic Tangent activation function was used to create the hidden layer with three nodes. Neural networks are very flexible models and tend to over-fit data. A Definitive Screening Design (DSD) was conducted to optimize the neural setting in order to minimize the over-fit concern. Both the Goodness Fit of Training set and Validation set can reach 99% R-Square. The Profiler Sensitivity analysis has shown that the Chocolate Type and Vitamin C are the most sensitive factors to predict the missed Cocoa%. The results also indicated that the "Fruit" Chocolate can be added as the 4th Chocolate Type. The Neural Black-Box algorithm revealed the hidden Chocolate Science and Product. This paper demonstrates the power of using the Engineering Design of Experiment (DOE) and Neural Network algorithm through STEAMS for the particular application of modeling chocolate products.

1 Introduction

Many people like eating chocolate but have concerns that chocolate is unhealthy. The objectives of this paper are: (1) clustering chocolate products, (2) performing missing value imputation, and (3) optimizing a neural network algorithm to deal with missing values. In order to accomplish these goals, the STEAMS Philosophy will be applied, which is presented

in the next sections. The main contribution of this paper is in illustrating how the STEAMS methodology for education can indeed teach students critical thinking while working on a real-world problem. Further, in this paper, several computing tools such as hierarchical clustering, dendrogram analysis and neural networks were employed. This paper is an extension of previous work by the authors [1] and provides additional information on the use of STEAMS in the educational experience.

1.1 The STEAMS Philosophy

The "STEAMS" (Science, Technology, Engineering, Artificial Intelligence, Mathematics, Statistics) methodology is a framework in which six components guide a developer (or, in our case, a student) through the design and analysis phase of a particular project. For this paper, (a) the *Science* is cocoa bean nutrition, flavonoids, flavanols, and antioxidants; (b) The *Technology* is the manufacturing process to produce the commercial chocolate products from coca beans; (c) the *Systematic Engineering* problem uses techniques to understand the root cause analysis. Further, the Design of Experiment is utilized to enhance the predictive modeling; (d) For the *Artificial Intelligence* component, several algorithms such as clustering and neural networks are utilized to recognize the patterns hidden among chocolate nutrition and products; (e) The *Mathematics* component uses such techniques as the dendrogram analysis tool to understand the clustering distance algorithms; And finally, the *Statistics* component uses graphical analytics to demonstrate the chocolate science and draw some practical conclusions. The JMP statistical software was used throughout this project. All six STEAMS elements are critical to making this project successful [1-4].

1.2 STEAMS Approach

A discussion of the science of the chocolate product is presented in [1]. This paper focuses on the computer software tools that may be applied to understand this science. In order to analyze the Chocolate Science and Nutrition pattern, the hierarchical clustering method is used for grouping similar nutrition variables into several representative clusters which are a linear combination of all variables in the same cluster. The cluster can be represented by the variables identified to be the most representative members of that cluster. The most representative variable in the cluster can be used to explain most of the variation in the data analyzed. The clustering method can effectively explore the chocolate nutrition clustering patterns which can explain the common foods science well. Adopting this dimension-reduction clustering algorithm can help simplify the predictive modeling by enhancing the signal-noise ratio, particularly in a very complicated/coupled design or system behavior. This formulates the basis of employing the STEAMS approach for analyzing the chocolate product.

2 Clustering and Neural Network of Missing Imputation

This section covers the following three subjects: (1) Data Collection, (2) the Clustering Method and Results, and (3) the Neural Network of Missing Value Imputation.

2.1 Data Collection

It's critical to collect the appropriate chocolate data in order to thoroughly analyze the chocolate product. A well-known commercial department store was chosen to collect the data, since it had plenty of chocolate products (a large enough sample size, as will be seen in the statistical analysis) and was extremely convenient for collecting data. Over sixty different types of chocolates were collected, and each had 20 nutrition variables. Not all 20 variables were used; instead, only 8 variables that were crucial to heart disease were used as shown in Figure 1.

Figure 1: Eight variables chosen to be used in statistical modeling based on scientific research

2.2 Hierarchical Clustering and Dendrogram

Hierarchical Clustering is a multivariate technique that groups chocolate products together that share similar values across a few nutrition variables. The method begins by treating each chocolate product as its own cluster. Then, at each step, the two chocolate products that are closest in terms of distance (features) are combined into a single cluster [5-12]. The result is depicted as a tree, called a Dendrogram, shown in Figure 2. The Dendrogram gives information about the degree of dissimilarity of clusters. Three clusters (Blue, Red, Green) were identified among 24 commercial chocolate products (no missing values). There are 39 other products missing Cocoa% information and which are not included in the Dendrogram Analysis.

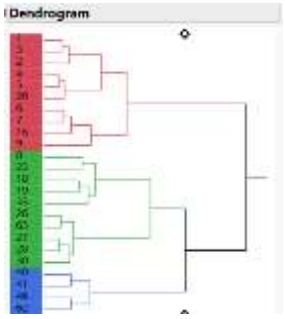
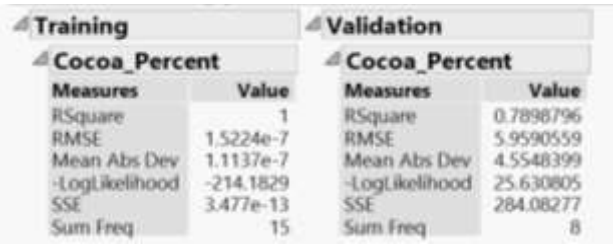


Figure 2: Hierarchical Clustering Dendrogram

2.3 Missing Value Imputation Algorithm

Among 63 commercial chocolate products, 39 products have missing Cocoa% information. In order to cluster these chocolate products (missing Cocoa%), a missing value imputation algorithm is commonly used. There are two missing value algorithms available in the JMP software: (1) Multivariate Normal Imputation and (2) Multivariate SVD Imputation [13-15]. The Multivariate Normal imputation method calculates pairwise co-variances to construct a covariance matrix for the response columns which in our case are the chocolate features. Then, each missing value is imputed by a method that is equivalent to regression prediction by including all the predictors that have no missing values. The Multivariate SVD imputation method avoids constructing a covariance matrix by using the well-known singular value decomposition. Both missing value imputation methods assume that there are no clusters, that the data come from a single multivariate normal distribution, and that the values are randomly missing. Further, both methods are based on Principle Component Eigen Analysis. Another approach to impute the missing values is through the use of neural networks, which has been popularly utilized in the Big Data Analytics [16-18]. A neural network uses a fully connected perceptron with one or more hidden layers. The functions applied at the nodes of the hidden layers are called activation functions. One of the most popular activation functions is the hyperbolic tangent (TanH) function available in JMP software which is used in this paper. Neural networks are very flexible models and tend to over-fit data. When that happens, the model predicts the fitted training data very well, but predicts future observations (validation data) relatively weaker. As shown in Figure 3, twenty-four available chocolate product nutrition data points available is split into 50%-50% Training Set-Validation Set for predicting the Cocoa%. Based on the JMP "Neural Network" analysis, the R-Square of Training Set is 1 (100%) and the R-Square of Validation Set is 0.79 (79%).



Training	
Cocoa_Percent	
Measures	Value
RSquare	1
RMSE	1.5224e-7
Mean Abs Dev	1.1137e-7
-LogLikelihood	-214.1829
SSE	3.477e-13
Sum Freq	15

Validation	
Cocoa_Percent	
Measures	Value
RSquare	0.7898796
RMSE	5.9590559
Mean Abs Dev	4.5548399
-LogLikelihood	25.630805
SSE	284.08277
Sum Freq	8

Figure 3: Model Goodness of Fit

The JMP Prediction Profiler can visualize and rank the top sensitive chocolate nutrition factors which can predict the Cocoa% as shown in Figure 4. The neural network results have indicated that (1) Chocolate Type, (2) Calcium Amount, and (3) Sugar Amount are the most sensitive factors to predict the Cocoa%. In general, among chocolate products, Dark Chocolate has higher Cocoa%, less Calcium than Milk Chocolate, and less sugar than Milk/White Chocolates [1]. It's not surprised that the Cocoa% can be predicted mainly by these top three parameters. It is interested to observe how the Neural Network method can help predict the missed Cocoa% in order to cluster all 63 Chocolate products.

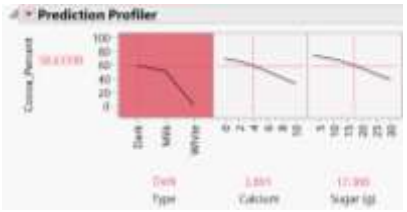


Figure 4: Prediction Sensitivity Profiler analysis [1]

How could the Neural Network algorithm predict the Cocoa%? As shown in Figure 5, with the output variable as Cocoa% and other Chocolate Nutrition as the input variables, a hidden layer of three nodes is added. Among the three hidden nodes, the third node has a higher sensitivity to predict the Cocoa%. The chocolate type has a higher sensitivity coefficient in the activation function from the nutrition layer to the hidden layer.

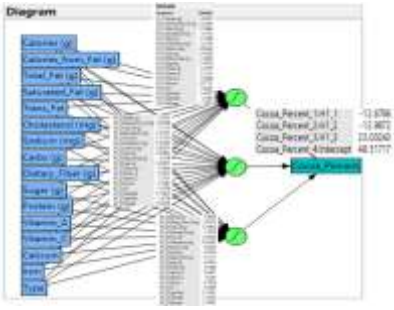


Figure 5: Neural Network Activation Functions [1]

Through the neural network activation functions, each missed Cocoa% can be predicted, based on the information on the other critical chocolate nutrition parameters. The constellation plots (shown in Figure 6) arrange the chocolate products as endpoints and each cluster join as a new point. The lines represent membership in a cluster. The length of a line between cluster joins approximates the distance between the clusters that were joined. Using the constellation plot, it is possible to see which clusters are combined first for different clustering algorithms. The dots with Green, Red and Blue colors are existing 24 Chocolate Products with full nutrition information while the black dots are products originally missed the Cocoa%. With the Neural Network algorithm of imputing the missing values, the entire 63 chocolate products can be clustered completely as shown in Figure 6.

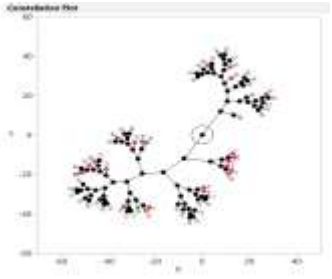


Figure 6: Constellation Plots of Neural Network Predictive Modeling

3 Optimize the Neural Network Algorithm of Missing Value Imputation

In Section 2, a neural network algorithm was utilized to impute missing value of Chocolate Cocoa% in order to complete the Hierarchical Clustering of chocolate products. In this section, in order to enhance the model predictability, we will further (1) perform a definitive screening design and design diagnostics, (2) optimize the neural network algorithm, and (3) optimize the relative importance of Goodness fit R-square between the training set and the validation set.

3.1 Definitive Screening Design and Design Diagnostics

Structured DSD DOE design was selected to conduct the Neural Network algorithm of missing value imputation. DSD design structure is near-orthogonal which could minimize the design confounding degrees while keeping the smaller DOE run size due to fold-over structure [18-22]. In the JMP Neural Network platform, there are three validation methods available as shown in Figure 7. The following two methods would be considered in optimizing the Neural algorithm:

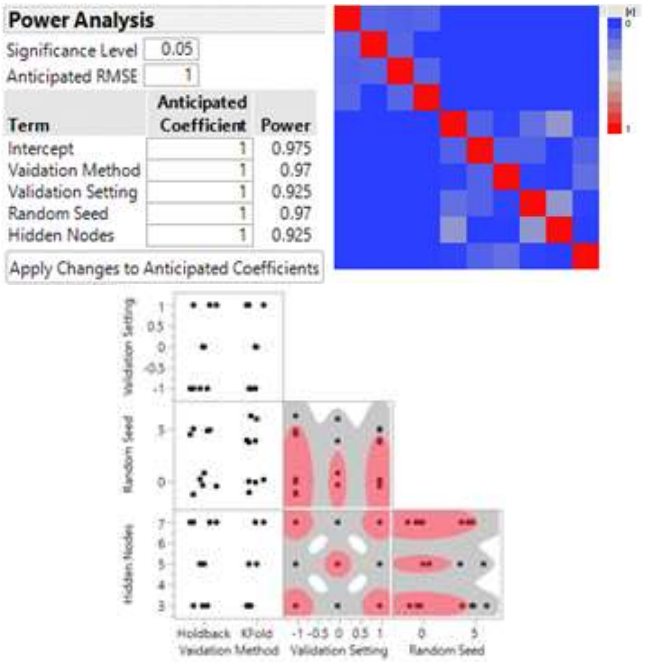


Figure 7: three design validation methods [1]

Two Design of Experiments (DOE) Output Goodness Fit Responses: (1) R-Square of Training Set, and (2) R-Square of Validation Set. An 18-Run definitive screening design (DSD) was conducted [1]. To ensure the DSD structure, three examination criteria was done before conducting the DSD simulation runs on the neural network algorithm.

The first power analysis is to check whether the sample size is enough. If the sample size is too small, the 95% confidence interval of any effect term will be very wide. Then, the power level would indicate the probability of the predicted sign is still valid. In Figure 7, all power levels are above 90% (no sample size concern). The second confounding color-map analysis is to investigate whether there are any Resolution II or Resolution III confounding concerns between any main effect and interaction effect. The confounding severity is indicated by a color map (from 0% correlation in blue to 100% correlation in red). The diagonal is always in red. In Figure 7, there is a very mild Resolution II confounding value (upper-left) due to the categorical variables and all blue in the Resolution III Zone (upper-right, and lower-left). Typically, we do not bother with the Resolution IV Zone (lower-right). Therefore, no confounding concern was noticed. The last uniformity analysis is to check whether the collected data is uniformly distributed in the design space (for continuous variables only). The denser area is shown in the red zone and less dense area is shown in the white zone. The white zone space should be minimized (in a symmetric way) to avoid the risk of the optimal design falling into the white zone (poor predictability). Figure 7 shows an acceptable symmetry and a small white zone. When the DSD optimization is done, we check whether the optimal design is falling into the white zone.

3.2 Neural Network Algorithm

A "definitive screening design" (DSD) would be conducted to optimize the "Neural" algorithm. Here are areas where definitive screening designs are superior to standard screening designs: (1) identify the causes of nonlinear effects by fielding each continuous factor at three levels and (2) avoid confounding between any effects up through the second order.

In DSD design, four Input Neural Algorithm Variables (available in JMP Neural Platform) are identified: (1) Validation Method (Holdback and K-Fold), (2) Validation Setting (Holdback portion and K) nested under the Validation Method. Nested DOE was used to address this "**Nested**" Validation Method and Setting limitation, (3) Computing Random Seed (Random of Fixed), (4) Number of Hidden Nodes.

Holdback: randomly divides the original data into the training and validation (holdback portion) sets. In Holdback menu, hold back portion can be input between o% and 100%.

K-Fold: divides the data into K subsets. Each K set used to validate the model fit on the rest of the data, fitting a total of K models. Chose model giving the best "**validation"** statistic. Best for small data sets (makes efficient use of limited data)

To further improve the Neural model prediction, the DSD Desirability Functions were set in Figure 8 Training Set, and Validation Set. The desirability functions are utilized to set the high, middle and low R-Square levels with assigned the desirability (i.e. how acceptable of the R-Square). The desirability range was set from 0.85 to 0.999 for both the Training and Validation. In order to overcome the Neural algorithm over-fit concern, the DSD optimization would favor the Validation by setting the higher importance at 2 over Training at 1.

Figure 8: desirability function of training set desirability function of validation set

3.3 DSD Result of Optimizing the Neural Network Algorithm

Section 3.3 will provide the DSD results of Section 3.2 DSD Optimization Plan. The objective of this DSD is to demonstrate the optimal Neural setting in order to improve the "over-fit" concern of lower model goodness fit on the validation set.

In Figure 9, the Optimal Neural Network Settings based on the desirability functions set are:

- "K-Fold" validation method is better than the "Holdback" method (due to small 24 sample size and favor in Validation portion)
- K=5 validation setting (24 available samples split into 5 subsets)
- Use Random Seed= 5 (fixed) to improve reproducibility over the Random Seed=0 (random)
- 4 Hidden Notes is best (avoiding over-fit due to limited 16 input variables for one hidden layer)

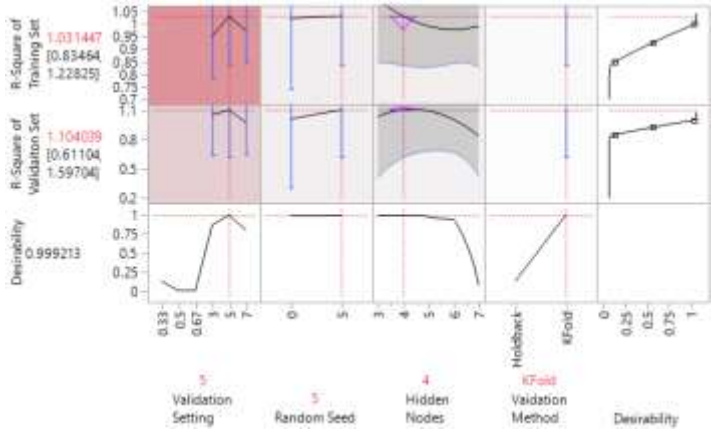


Figure 9: DSD Profiler of Neural Setting [1]

Both R-Squares are over 1 (100%) which may indicate the regression residual distribution is not in Normal Distribution (lack of fit concern). Also, the confidence interval of R-Square is +/- 0.2 for the training set and +/-0.5 for the validation set. Such poor modeling uncertainty may be due to the smaller sample size available and which will limit the reliability of the number of hidden nodes (details of the wide confidence interval are provided in [1]). Though,

the overall optimization desirability is almost 1. Authors won't address these concerns in this paper, additional effort is conducting in order to further understand the true mechanisms. One approach is to collect more Chocolate data to increase the sample size and reduce the confidence interval. The other approach is to examine each existing Chocolate data quality and to filter out some outliers.

The optimal Neural setting was further validated based on the 24 available Chocolate products and details are provided in [1]. A summary is shown in Figure 10. Both the R-Square of Training and Validation are beyond 0.99. The optimal neural setting has significantly improved the validation goodness fit R-Square by more than 0.2 (20%) while sustaining the very perfect training R-Square. The sensitivity ranking of predicting the Cocoa% are slightly different from the previous one in Figure 5. Chocolate Type is still the top factor but not a dominant one anymore. Vitamin_C is emerged as second choice as a surprise. When we revisit the Chocolate Products, several Fruit Chocolate Products are showing higher Vitamin_C which was not paid full attention during the defining the Chocolate Type. It may suggest that, in addition to Dark, Milk, White chocolates, the fourth "Fruit" chocolate shall be added. The optimal Neural Algorithm is just enhanced the modeling goodness fit, but also reveal the hidden Chocolate Science and Product in the Profiler Sensitivity Analysis.

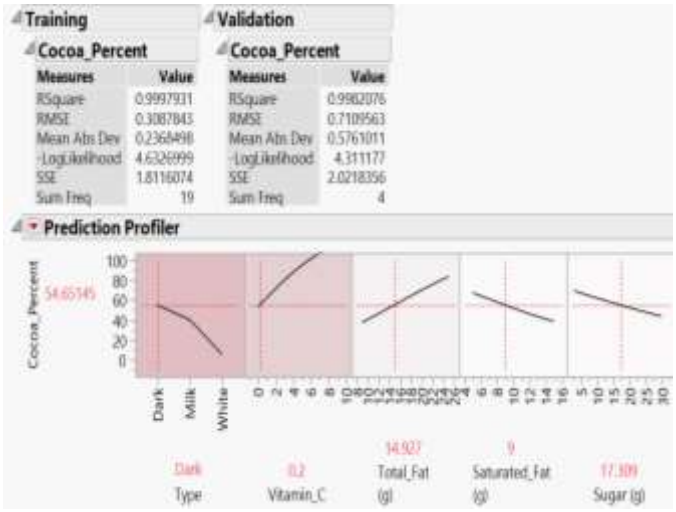


Figure 10: validate the optimal Neural setting [1]

Most criticism on Neural algorithm is "Black Box" transformation. However, most data scientists shall dig deeper on the modeling sensitivity and map these patterns to the real "Science and Engineering" world. This is the main vision and guiding principle of this paper to emphasize the "STEAMS" approach.

3.4 Optimize Desirability Importance Factor

2-Factor & 3-Level Full Factorial DOE designs were constructed as an approach to improve the Optimal Design Desirability as well as mitigate the risk of model-overfitting. The table displayed per Figure 11 provides an examination of the relative importance level of the R-Square between the training and validation set on a binary scale with integers 1 and 2. When the importance level is equal to 1, the optimization algorithm follows a linear regression curve in order to select the optimized Neural design. When the importance level is equal to 2, the optimal design shall be chosen based on the curve defined at Power = 2. The opportunity framework of this optimization scenario imposes a larger penalty if the calculated R-Square does not meet the desired R-Square level at Power = 1 as compared at Power = 2. We also established a higher threshold for Desirability R-Square (Upper Limit = 0.999, Target = 0.95, Lower Limit = 0.9) in an effort to achieve a higher Goodness-of-Fit (GoF) for each model iteration. We reemphasize that in this particular section, the primary objective is to further optimize the importance level of two R-Squared values, in the Training set and the Validation set, respectively, in order to mitigate the risk of model-overfitting (given that Neural algorithms are well-known to be susceptible to overfitting). Higher performing models all demonstrate a similar trend: the presence of 3 hidden nodes which are mainly associated with the training set and 4 hidden nodes which are mainly associated with the validation set. The optimal setting for the Neural model is still the same as the one shown previously (Figure 9), which makes sense because in this DOE structure, all model desirability is already maximized at values above 0.999. Nevertheless, a structure of (Relative) Importance Settings and Desirability indices may prove to be an innovative approach in other experimental situations (raising the bar for expectation on R-Squared).

Importance Setting		Optimization Result	Optimal Neural Settings			
Importance of Training Set	Importance of Validation Set	Overall Desirability	Validation Method	Validation Setting	Random Seed	Hidden Nodes
1	1	0.9997	KFold	5	5	3
1	1.5	0.9997	KFold	5	5	4
1	2	0.9997	KFold	5	5	4
1.5	1	0.9997	KFold	5	5	3
1.5	1.5	0.9997	KFold	5	5	3
1.5	2	0.9997	KFold	5	5	4
2	1	0.9997	KFold	5	5	3
2	1.5	0.9997	KFold	5	5	3
2	2	0.9997	KFold	5	5	3

Figure 11: full factorial DOE of optimizing the importance settings

Constellation plots are depicted in Figure 12 and demonstrate the risk versus benefit tradeoff associated with optimization of the Neural Settings from the original default and DSD Optimization compared to the corresponding Importance level. The Constellation plots reveal that the optimal Neural setting has significantly impacted the clustering process, suggesting that the mis-classification risks of grouping clusters can effectively be minimized.

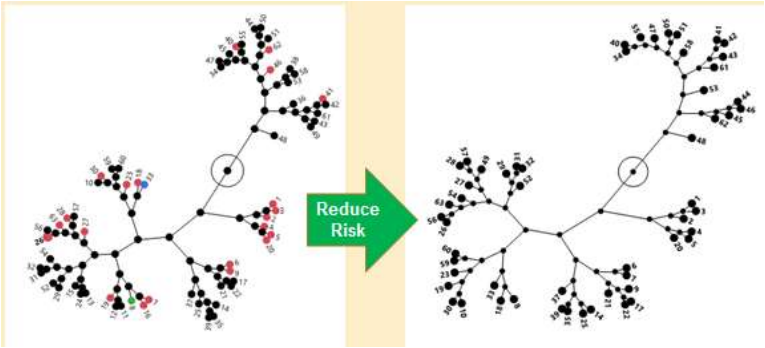


Figure 12: constellation plots of default Neural Setting and optimal Neural Setting

4 Conclusions

The STEAMS approach is very successful on understanding Chocolate Science Research and Nutrition Food Science. Modern Multivariate Clustering Statistics and Artificial Intelligence Neural Network Algorithms can explore the Chocolate Science Patterns which can further help consumers pick their healthy chocolate products based on their preferred nutritions needed. Neural Network algorithm can help missing value imputation and enhance the clustering method. The classical DSD can further help optimize the Neural setting in order to enhance the capability of discovering the hidden patterns. This STEAMS approach can be applied to similar fields such as Coffee Science and Product, as well as to other Healthy Nutrition Study.

Acknowledgement

Authors would like to thank the Biostatistics Advisor Patrick Giuliano and CATA 2020 Program Co-Chair Gordon Lee for their paper reviewing and suggestions.

References

1. Chen, Mason, “Methodology of Conducting Scientific Research", 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019
2. Chen, Mason, “Multivariate Statistics of Antioxidant Chocolate”, IWSM Bristol Proceedings, Vol 2 37-40, 2018
3. Chen, Mason, “Choose Healthy Chocolate”, IEOM Europe Proceedings, 434-441, 2018
4. Wu, Anna, “Starbucks and Cardiovascular Disease Prevention.” IEOM, IEOM Society, 2018
5. Harris, C.W. and Kaiser, H.F., “Oblique Factor Analytic Solutions by Orthogonal,1964

6. Milligan, G.W., “An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms,” Psychometrika, 45, 325–342. 1980

7. Hartigan, J.A., “Consistence of Single Linkage for High–Density Clusters,” Journal of the American Statistical Association, 76, 388–394,1981

8. Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S., “Sur La Liaison et la Division des Points d’un Ensemble Fini,” Colloquium Mathematica, 2, 282–285, 1951

9. Jardine, N. and Sibson, R., Mathematical Taxonomy, New York: John Wiley and Sons,1971

10. McQuitty, L.L., “Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies,” Educational and Psychological Measurement, 17, 207–229, 1957

11. Sokal, R.R. and Michener, C.D., “A Statistical Method for Evaluating Systematic Relationships,” University of Kansas Science Bulletin, 38, 1409–1438, 1958

12. Sneath, P.H.A. “The Application of Computers to Taxonomy,” Journal of General Microbiology,17, 201–226, 1957

13. Golub, G.H., Kahan, W., “Calculating the singular values and pseudo-inverse of a matrix,” Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis 2:2, 205–224., 1965

14. Golub, G.H. and van der Vorst, H.A., “Eigenvalue Computation in the 20th Century,” Journal of Computational and Applied Mathematics 123, 35-65., 2000

15. Press, W.H, Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes in C: The Art of Scientific Computing, Second Edition, Cambridge, England: Cambridge University Press, 1988

16. Schmidhuber, J., "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117 , 2015

17. Ruslan, Salakhutdinov, Joshua, Tenenbaum "Learning with Hierarchical-Deep Models". IEEE Transactions on Pattern Analysis and Machine Intelligence. **35** (8): 1958–71,2012

18. Errore, A., Jones, B., Li, W., and Nachtsheim, C., “Using Definitive Screening Designs to Identify Active First- and Second-Order Factor Effects,” forthcoming, Journal of Quality Technology, 2016

19. Jones, B. and Nachtsheim, C.J., “Efficient Designs with Minimal Aliasing,” Technometrics, 53:1, 62-71,2011

20. Jones, B. and Nachtsheim, C.J., “Definitive Screening Designs with Added Two-Level Categorical Factors,” Journal of Quality Technology, 45, 121-129,2013

21. Jones, B. and Nachtsheim, C.J., “Blocking Schemes for Definitive Screening Designs,” Technometrics, 58:1, 74-83, 2016

22. Miller, A. and Sitter, R. R., “Using Folded-Over Nonorthogonal Designs,” Technometrics, 47(4), 502-513,2005Mason Chen. "Introduce a Novel "STEAMS"