Neural Networks for Tea Leaf Classification

Retraction: Neural Networks for Tea Leaf Classification (*Journal of Physics: Conference Series* **1432** 012075)

Jesús Silva[1], Hugo Hernández Palma[2], William Niebles Núñez[3], Alex Ruiz-Lazaro[4] and Noel Varela[5]

[1] Universidad Peruana de Ciencias Aplicadas, Lima, Perú.

[2] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.

[3] Universidad de Sucre, Sincelejo, Sucre, Colombia.

[4] Universidad Simón Bolívar, Barranquilla, Atlántico, Colombia

[5] Universidad de la Costa, Barranquilla, Atlántico, Colombia

Published 15 September 2020

This article, and others within this volume, has been retracted by IOP Publishing following clear evidence of plagiarism and citation manipulation.

This work was originally published in Spanish (1) and has been translated and published without permission or acknowledgement to the original authors. IOP Publishing Limited has discovered other papers within this volume that have been subjected to the same treatment. This is scientific misconduct.

Misconduct investigations are ongoing at the author's institutions. IOP Publishing Limited will update this notice if required once those investigations have concluded.

IOP Publishing Limited request any citations to this article be redirected to the original work (1).

Anyone with any information regarding these papers is requested to contact conferenceseries@ioppublishing.org.

(1) Paula, Luisina A. de | Eckert, Karina B. | Guismín, Gabriel (2018), Clasificación de hojas de té al ingreso del proceso de secado mediante redes neuronales con datos supervisados y no supervisados, X Congreso de AgroInformática (CAI) - JAIIO 47 (CABA, 2018) (http://sedici.unlp.edu.ar/handle/10915/70998)

Neural Networks for Tea Leaf Classification

Jesús Silva[1], Hugo Hernández Palma[2], William Niebles Núñez[3], Alex Ruiz-Lazaro[4] and Noel Varela[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[3]Universidad de Sucre, Sincelejo, Sucre, Colombia.
[4]Universidad Simón Bolívar, Barranquilla, Atlántico, Colombia
[5]Universidad de la Costa, Barranquilla, Atlántico, Colombia

[1]Email: jesussilvaUPC@gmail.com

Abstract. The process of classification of the raw material, is one of the most important procedures in any tea dryer, being responsible for ensuring a good quality of the final product. Currently, this process in most tea processing companies is usually handled by an expert, who performs the work manually and at his own discretion, which has a number of associated drawbacks. In this work, a solution is proposed that includes the planting, design, development and testing of a prototype that is able to correctly classify photographs corresponding to samples of raw material arrived at a dryer, using intelligence techniques (IA) type supervised for Classification by Artificial Neural Networks and not supervised with K-means Grouping for class preparation. The prototype performed well and is a reliable tool for classifying the raw material slammed into tea dryers.

1. Introduction

The process of classification of the raw material that above the dryers is responsible for determining the percentages of weeds in the shipments, separating them from what can be considered as a processable raw material, which at the same time is classified in 3 types known as green shoots, green stems and non-green leaves or stems respectively. Currently, both processes are done by hand, from a small sample taken from each shipment that is above the processing plant. For this process, an expert is needed who can visually detect the differences between each type corresponding to the classification, since the classification decisions of the classification will depend on the future production quality [1] and the obtaining of raw material proper steps in the productive [2] process. It should be noted that in the event of the absence (momentary or indeterminate) of this specialized person, it would be difficult to replace it quickly and classify it in the same way, thus maintaining the uniform quality of the product [3][4].

Because of this situation, there is a need for a computer solution that allows a uniform, fast and effective classification of the raw material reached to tea dryers. The solution to this problem involves, first of all, the digitization of images and then the application of data processing techniques to perform such classification.

To accomplish this, the images can be represented numerically by using their RGB (Red-Green-Blue) properties of each of their pixels and different mathematical descriptors that allow them to be specified. From numerical representation, AI techniques, in this scenario, are an appropriate option. This

is because they have the ability to make different types of classifications, especially with regard to image recognition, such as that carried out in this research [5].

To verify the applicability of this process it is necessary to formalize it and carry out a series of conclusive tests. To address the problem posed, Artificial Neural Networks have been chosen as an artificial intelligence technique to carry out the relevant classifications and the K-means clustering technique, for obtaining data unsupervised, thus achieving measuring the fitness of the qualifying technique.

2. Materials and Methods

2.1 Grouping

Clustering is the process of organizing instances of data into groups whose members are similar in some way. A cluster is a collection of data instances that are similar to each other and not similar to data instances in other clusters (groups). The task of clustering is to find the groups hidden between the data. For humans, it's not always easy to visualize or detect a cluster or classification groups, so clustering is a very good tool to help with this task [6]

Clustering is considered as a type of unsupervised learning; This is because, unlike supervised learning, class values do not denote an apriori partition or grouping of data, but instances or classes are discovered throughout the partition task [7].

One of the most popular grouping algorithms is K-means, which achieves given a certain number of points and the number k of clusters (specified by the user), iteratively partitioning the information into k clusters based on a distance function. It consists of randomly selecting k points as seed centroids that will then be used to perform distance calculations between each centroid and each existing point in the data series. Each point is assigned to the centroid that is closest to itself. Once all points are assigned, the centroid of each cluster is recalculated. The process is repeated until a stop criterion [7] is known or carried out.

2.2 Artificial Neural Networks

"Neural networks are sets of simple, usually adaptive calculation elements, massively interconnected in parallel and with a hierarchical organization that allows you to interact with some system in the same way that the nervous system does biologically." [8].

Because of their constitution and fundamentals, RFs have a large number of brain-like characteristics. For example, they are able to learn from experience, to generalize from previous cases to new cases, to abstract essential features from entries that represent irrelevant information, etc. This means that they offer numerous advantages and that this type of technology is being applied in multiple areas. Some advantages and main features of neural networks are [9]:

- Adaptive learning: Ability to learn how to perform tasks based on a workout or initial experience.
- Self-organization: A neural network can create its own organization or representation of the information it receives through a learning stage.
- Fault Tolerance: Partial destruction of a network leads to a degradation of its structure; However, some network capabilities can be retained, even suffering great damage.
- Real-time operation: Neural computations can be performed in parallel; this is what machines with special hardware are designed and manufactured to obtain this capability.
- Easy insertion within existing technology: Specialized chips can be obtained for neural networks that improve their ability in certain tasks. This facilitates modular integration into existing systems.

2.3 Neural Networks with Backward Propagation (BPN)

Neural Networks with Backward Propagation (BPN) use a supervised learning algorithm. Given a pattern to the network input as a stimulus, it propagates forward from the first layer through the next layers of the network, generating an output, it is compared to a desired output (which was stipulated) and an error signal is calculated for each of the themselves obtained. The error outputs are propagated

backwards, starting from the output layer, to all neurons in the hidden layer that contribute directly to it. It should be noted that hidden layer neurons only receive a fraction of the total error signal, based approximately on the relative contribution that each neuron has contributed to the original output. This process is repeated, layer by layer, until all neurons in the network have received an error signal describing their contribution relative to the total error. [10]

The importance of BPNs is their ability to learn the relationship between a set of patterns and their corresponding outputs in order to apply that same relationship after training to noisy patterns where an output is activated if the new input is similar to the presented in learning, which makes BPN systems capable of generalizing [9].

3. Results

The prototype developed was implemented in the Java programming language, with the feature of having a mostly proprietary graphical user interface and coding, as well as the reuse of small portions of code belonging to different authors [11] and bookstores [12].

3.1 Obtaining the images

The photo samples were taken twice during the months of November to March of the 2016-2017 season of tea; these were taken in the settlements of valle del Cauca Colombia, in broad daylight (morning hours) and at the work bench used precisely for the manual classification process of the raw material arrived.

The photographs were taken from above, in a straight line and at a distance of approximately 50 cm, allowing to be observed the white background used on the workbench, as well as the characteristics of each of the samples.

In total, 630 photographs were taken, with a 14.1 megapixel Sony W320 digital camera; of these, 420 were chosen according to their conditions to be processed (this refers to them having good quality, not being moved, in low light or other common defects in the photographs); at the same time, 210 of the latter were selected that have a typical representation of the raw material with which they are usually worked, as well as strange or different situations that could be given in the dryer, thus analyzing the points extremes (or limit situations) that could happen.

Of the last 210 photographs cited, 90 of them were taken for the training of artificial intelligence techniques (30 shoots, 30 green stems and 30 stems and non-green leaves, in equal proportion and indicated correspondingly by the expert responsible for the classification tasks) and the remaining 120 for the relevant tests, where 60 are intended to be for "perfect" situations (20 of pure shoots, 20 of pure green stems and 20 of pure stems and leaves), 30 for "common" situations (where The green shoot abounds) and 30 for "strange" situations (10 for foreign objects, 10 where stems abound and 10 where non-green leaves abound).

It should be noted that while the photographs were taken with a camera of 14.1 mega pixels (4320x3240 pixels), these were adapted to a size of 256x256 pixels (without loss of quality or any properties) to allow a lot of processing of the images faster and faster, reducing potential waiting times.

3.2 Image descriptors used

In order to use the photographs obtained, it was necessary to find a way to represent them numerically, so that they can be used as data of the training and testing processes themselves. For this, the RGB properties of each of the images were used; these indicate the characteristics of colors in red, green and blue (Red, Green, Blue) numerically for each pixel of a photograph; [13] that is, if each photograph has a size of 256x256 pixels you get a total of 65,536 pixels which, in turn, have 3 color values each, so the prototype works with a total of 196,608 data per photograph.

Once the prototype achieves the numerical characteristics of each photograph in RGB, in order to summarize in a certain way the captured information, we proceeded to use different mathematical descriptors: The mean, the median, the variance, the standard deviation, kurtosis coefficient and linear correlation [14].

3.3 Configurations for testing

Both monitored data (provided by the prototype from expert user instructions) and unsupervised (provided by the unsupervised k-Means clustering technique) have been used to perform the different tests.

The numerical values that represent each of the photographs for training and testing depend on the settings cited in Table 1.

Table. 1. Configurations to use for planned tests

Configuration	Features
1	**Descriptors**: Mean, median, standard deviation, variance, kurtosis coefficient and linear correlation. **RNA**: 8 descriptors x 2 values (RGB) - 16 input neurons
2	**Descriptors**: mean, median, and standard deviation. **RNA: 4** descriptors x 2 values (RGB) x 8 input neurons.
3	**Descriptors**: mean, variance and Curtosis coefficient. **RNA: 4** descriptors x 3 values (RGB) x 12 input neurons.

3 (three) test sets have been conducted. The first using pure photographs of the samples (shoot, stem, non-green leaf), the second, using photographs reflecting a fairly common situation in the dryer, which occurs when there are mostly shoots in the sample and the use photographs with strange situations that may occur (samples with a majority number of stems, non-green leaves or the presence of foreign objects). This can be seen in a summary way in Table 2.

Table. 2. Tests performed

Test	Features
1	**Perfect Situations**: 22 Photos of Outbreaks, 22 Stem photos, 22 Photos of Non-Green Leaves
2	**Common Situations:** 34 photos of Mostly Outbreaks
3	**Strange Situations:** 12 photos of Mostly Stems, 12 Photos of Mostly Non-Green Sheets, 12 Photos of Strange Objects

3.4 Classification tests with Neural Networks and supervised data

Testing using the RNA artificial intelligence technique was carried out with the prototype developed, achieving the classifications that can be observed in a summary way in Tables 3, 4 and 5 for those who have used supervised data.

Table. 3. Test 1 results with RNA and monitored data

Configuration	Correct Classification	Incorrect classification	% hit
1	62	0	100 %
2	62	0	100%
3	60	1	97,23%

Table. 4. Test 2 results with RNA and monitored data

Configuration	Correct Classification	Incorrect classification	% hit
1	33	0	100%
2	33	1	96,67%
3	30	0	100%

Table 5. Test results 3 with RNA and monitored data

Configuration	Correct Classification	Incorrect classification	% hit
1	26	4	87,34%
2	29	1	96,67%
3	27	3	83,34%

The results of test No. 1 are presented in Table 3. As can be seen, the results obtained with these configurations yielded rating percentages greater than or equal to 97.23%, among which are configurations 1 and 2 that obtained 100% success.

Table 4 shows the summary of the results obtained in Test No. 2. This test reflects one of the most common situations in the scenario of a tea dryer, where most of the content of the sample above represents tea shoots with a small possible presence of other parts of the raw material. As can be seen, as in test No. 1, the hit percentages exceed or equal 96.67%.

In Table 5, the results obtained from Test No. 3 can be reflected. This evidence shows strange situations that could occur in a tea dryer to see the prototype's reaction to the presentation of boundary situations, such as the presence of cans or foreign objects, more other parts of the raw material that are not outbreaks, etc. When training the prototype with pure shoots, stems and non-green leaves and then testing with samples of relevant foreign situations, it can be visualized that in all configurations hit percentages equal to or greater than 83.34 have been obtained %, there are mostly problems in detecting situations where green stems and non-green leaves abound; however, it can be noted, when looking in detail at each of the results obtained, there are virtually no problems in the recognition of foreign objects present in the samples, which is the worst situation that can occur and is of great importance that the prototype was capable of correctly classifying it [15].

3.5 Classification tests with Neural Networks and Unsupervised Data

The results obtained in the classification tests carried out with Neural Networks and unsupervised data are reflected and sorted in Tables 6, 7 and 8; each of these were caused by the use of training data obtained in an unsupervised manner through the K-means clustering technique, by the prototype developed, which specifies the number of classes in which the aforementioned should group the data (in this case 3 samples).

Table. 6. Test 1 results with RNA and unsupervised data

Configuration	Correct Classification	Incorrect classification	% hit
1	58	4	96 %
2	55	6	89,34 %
3	55	6	89,34 %

Table. 7. Test results No. 2 with RNA and unsupervised data

Configuration	Correct Classification	Incorrect classification	% hit
1	30	2	97,672%
2	33	0	100 %
3	6	22	15,89 %

Table. 8. Test results No. 3 with RNA and unsupervised data

Configuration	Correct Classification	Incorrect classification	% hit
1	22	12	70,68 %
2	12	18	34,34 %
3	18	16	58,63 %

Table 6 presents the results of the No. 1 test performed with the different configurations and with the use of the data created in a non-supervised way by K-Means clustering. As can be seen, with all three configurations, hit percentages greater than 89.34% were obtained; however, the results are not as high as those manifested with samples obtained in a supervised manner. This is undoubtedly due to the way Clustering groups the samples to be used, as there is some possible degree of error when assembling classes in an unsupervised manner.

The results of test No. 2 are presented in Table 7. On the one hand, configuration No. 2 has provided a perfect classification, even though the data has been configured in an unsupervised manner, as has configuration No. 1 that presents only an error when performing the classification; However, the no. 3 configuration yields a really low value that is 15.89%. This is due to the combination of data obtained in a non-clustered manner along with the use of different mathematical descriptors (which do not match configurations 1 and 2).

Table 8 presents the results of the No. 3 test performed with the different configurations mentioned above and with the use of data created in a non-clustered manner. When working with unsupervised data and strange situations that do not often occur in the dryer, it can be observed that all results have success rates below 70.68%.

3.6 Final Test Results

In Table 9, you can see the different denominations of the classifications made according to their success rate; those that have provided better results, are classified as good, very good or perfect, thus allowing to corroborate which combination of values is best suited to perform classifications using neural networks and samples Presented.

Table. 9. References of the names to the classifications according to their percentage of success.

Denomination	Range
Perfect	100%
Very good	90% - 99,99%
Good	80% - 89,99%
Regular	50% - 79,99%
Bad	20% - 49,99%
Very bad	0% - 19,99%

4. Conclusions

Based on the tests carried out with the prototype and the different configurations, those that provided the highest amount of success percentage for each of the tests were chosen. From this, it can be said that the classification Artificial Intelligence (AI) technique studied in this work (RNA) is suitable for the classification of raw material entered into a tea dryer.

RNA achieved very good results in each of the tests, both with monitored and unsupervised data (generated by Clustering K-means), with the exception of the third of them, where the samples used were intended to generate a different or little reaction usual in the prototype, since they coincided little and nothing with the images used for training, (belonging to bud, stem and non-green leaf in pure and unaltered state).

Regarding the generation of the data and samples, supervised data were used, which, for its acquisition, demand the presence of an expert human user to organize the corresponding initial samples, indicating which of them belong to each of the classes. This issue, takes an extra time that includes the organization and preparation of test samples that adapt to the scenarios raised, so that once the technique is trained and configured, it is based on these criteria to perform the classifications Corresponding.

On the other hand, it was proposed, to obtain data in an unsupervised way, which does not require an organization of the samples or specifications of each of the types, since using the K-means clustering technique, the designed prototype is responsible for organizing the photographic samples presented to

you according to the similarity criteria found in the RGB values of each of them (represented by mathematical descriptors), considerably reducing the time involved in classification.

References

[1] Ministerio de Ciencia, Tecnología e Innovación Productiva, Profecyt, Agencia de promoción Científica y Tecnológica, Unión Industrial Argentina. Té en Misiones: Debilidades y Desafíos tecnológicos del sector productivo. Buenos Aires, Argentina (2012).
[2] Zamora, K., Castro, L., Wang, A., Arauz, L. F., & Uribe, L. (2017). Potential use of vermicompost leachates and tea in the control of the American leaf spot of coffee Mycena citricolor. Agronomía Costarricense, 41(1), 33-51.
[3] Jain, Mugdha, and Chakradhar Verma. "Adapting k-means for Clustering in Big Data." International Journal of Computer Applications 101.1 (2014): 19-24.
[4] Hariri S, and M. Parashar.Tools and Enviroments for Parallel and Distributed Computing. John Wiley & Sons. ISBN 0-471-33288-7, pag 229, 2014.
[5] Viloria, A. "Commercial strategies providers pharmaceutical chains for logistics cost reduction." Indian Journal of Science and Technology 8, no. 1 (2016).
[6] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371.
[7] N. Sapankevych y R. Sankar, “Time Series Prediction Using Support Vector Machines: A Survey”, IEEE Computational Intelligence Magazine, vol. 4, núm. 2, pp. 24–38, may 2009.
[8] F. Villada, N. Muñoz, y E. García, Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, Información tecnológica, vol. 23, núm. 4, pp. 11–20. 2012.
[9] Venugopal K, K.G. Srinivasa and L. M. Patnaik. Soft Computing for Data Mining Applications. Springer Berlin Heidelberg: Springer-Verlag. ISBN 978-3-642-00192-5, pp 354, 2009.
[10] Brdar S., Culibrk D., Marinkovic B., Crnobarac J., Crnojevic V. Support Vector Machines with Features Contribution Analysis for Agricultural Yield Prediction, Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment, 43-47, 2011
[11] Choudhury, A. and Jones, J. Crop yield prediction using time series models, Journal of Economics and Economic Education Research., 15, 53-68, 2014.
[12] R. Putha, L. Quadrifoglio, and E. Zechman. Comparing ant colony optimization and genetic algorithm approaches for solving traffic signal coordination under oversaturation conditions. Computer- Aided Civil and Infrastructure Engineering, 27(1), 14-28, 2012.
[13] D. Teodorović, and M. Dell’Orco. Mitigating traffic congestion: solving the ride-matching problem by bee colony optimization. Transportation Planning and Technology, 31(2), 135-152, 2008.
[14] Amelec, V., & Alexander, P. (2015). Improvements in the automatic distribution process of finished product for pet food category in multinational company. Advanced Science Letters, 21(5), 1419-1421.
[15] Salinas-Piélago, J. (2016). Revisión sobre el uso del mate de hoja de coca en la prevención del mal agudo de montaña. Revista de Neuro-Psiquiatría, 79(3), 166-168.