Trends in Technical & Scientific Research
ISSN: 2641-8355

Juniper Publishers
key to the Researchers

Research Article
Volume 4 Issue 4 - December 2020
DOI: 10.19080/TTSR.2020.04.555643

Trends Tech Sci Res

Copyright © All rights are reserved by Trung Pham

Optimization of Raw Material Yield Using Data Mining

Trung Pham[1]*, Teresita Hernández Toledo[2] and Karla Moraga Correa[2]

[1]*Information Technology Research Center, University of Talca, Chile*

[2]*School of Business Informatics Engineering, University of Talca, Chile*

Submission: November 3, 2020; **Published:** December 03, 2020

***Corresponding author:** Trung Pham, Information Technology Research Center, University of Talca, Chile

Abstract

This project uses data mining to group the operational data of the dehydration process of the Granny Smith apple at the Agroindustrial Surfrut company so that an optimal point in the operation can be identified. This approach is selected because the operational data is so complex that it is not possible to model it in a mathematical expression necessary to identify an optimal point through mathematically solving an optimization problem. The optimal point identified in the data mining will be used by the company Agroindustrial Surfrut to lower operating costs while maintaining the quality of its products, allowing a competitive advantage in the national market.

Keywords: Data mining; K-means method; Dehydration process; Statistical method

Introduction

A food production process [1] is a process that takes edible material and prepares it to a point of sale to consumers. Preparation steps can usually be tailored according to the type of food, the salable point of preparation, consumer expectations, consumer demand, etc. Specifically, in the context of this work, the dehydration process [2] for the Granny Smith apple [3] is considered, where a flow of hot air evaporates and removes the water content in the fruit. The hot air temperature will scale the temperature of the fruit and evaporate its water content into air, and the air flow will remove this water vapor. These variables are adjusted to achieve the quality of the final dried fruit product. An important factor that must be considered to operate this process is the energy consumption that proportionally affects the cost of the operation.

In a company, the operating cost must be optimized [4] in a common routine to increase its competitive advantage [5] through better work efficiency [6] , better profit margin [7] , low price of the products offered, etc. In dehydrating apple fruit, the use of energy to scale the air temperature and move the hot air flow significantly contributes to the cost of the operation. Therefore, this energy cost must be minimized in a constraint [8,9] to maintain the good quality of the final product (dried apple). If a company can minimize its cost of operation, the Agroindustrial Surfrut company, the sponsor of this project, which produces dried apple through the dehydration process, already maintains a set of historical data during its existence. Due to the fact that there is no mathematical model relating precisely the variables independent and observable with the quality of the final product and energy to solve an optimization problem that maximize the quality of the dried fruit while that minimize energy consumption, The use of data mining is proposed to discover the values of the independent and observable variables that provide the optimality [10,11] of the operation of the dehydration process of the Granny Smith apple.

In this work, the data mining process [12,13] is applied to a large data set in many dimensions, with each dimension representing an observable variable of the dehydration process. This set is divided into multiple clusters through the K-means method [14,15] based s on the similarity of the data in the data grouping step. Each cluster is analyzed to extract a set of parameters that represents the cluster. All the clusters are compared with each other to determine the cluster that delivers the best performance in terms of quality dried fruit and energy consumption. The parameters that this cluster represents are considered the optimal values for the independent and observable variables of the dehydration process.

Background

In the optimization of a process, said process must be modeled in a mathematical expression representing it. This expression is formulated in an objective function that must be optimized in the sense that a minimum point of the objective function is found. Figure 1 shows three steps in this process with an arbitrary three-dimensional example. In this example, the x and y dimensions represent the independent variables, and the z dimension represents the cost dependent variable. C hen data is formable a function analytic is constructed in the form of a mathematical expression as a model representing the data. With this model, an optimal point can be calculated analytically in a systematic and precise way. However, data is not always modellable, and the approach of modeling it with mathematical expression cannot always be used.

In the first step of Figure 1, data is collected for various scenarios during the operation of a process. When these data can sufficiently represent the process, they can be modeled in a mathematical expression with calculated parameters based on minimizing the error between the mathematical expression and the collected data. This minimization is formulated in an error optimization problem where the error is minimized without constraint as in the following:

$$\min_{a1,a2,...aN} \sum_x \sum_y \sum_z \| f(x,y,z,a1,a2,....aN) - \begin{bmatrix} x \\ y \\ z \end{bmatrix} \|^2 \qquad (1)$$

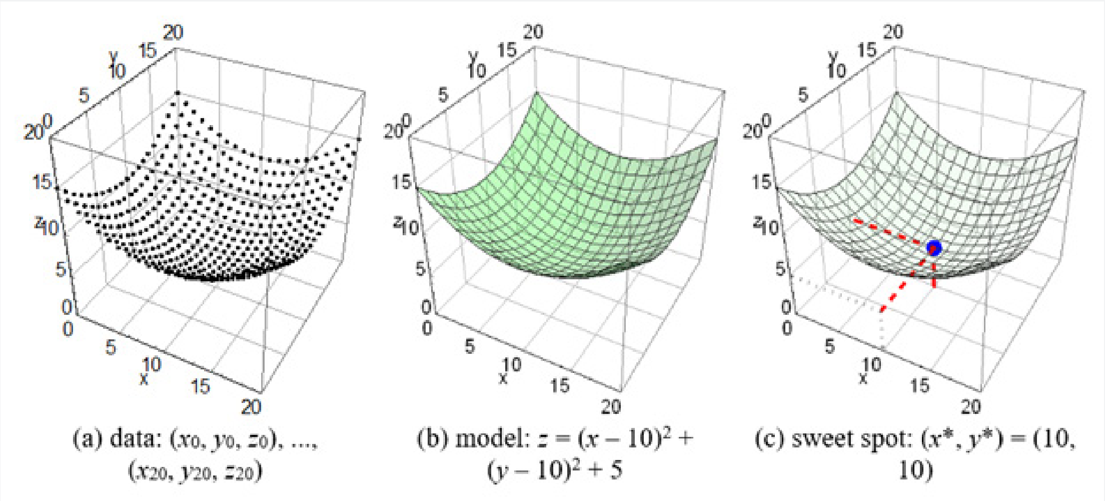
where a1, a2, ..., a N are parameters and x, y, z are variables of the function f (·), and the operator || · || is the norm operator that calculates the magnitude of a vector. Minimize the objective function in (1) is a routine task ria to solve a system of equations

$$\frac{\partial}{\partial a_2} \sum_x \sum_y \sum_z \| f(x,y,z,a_1,a_2,...,a_N) - \begin{bmatrix} x \\ y \\ z \end{bmatrix} \|^2 = 0 \qquad (2a)$$

$$\frac{\partial}{\partial a_2} \sum_x \sum_y \sum_z \| f(x,y,z,a_1,a_2,...a_N) - \begin{bmatrix} x \\ y \\ z \end{bmatrix} \|^2 = 0, \qquad (2b)$$

$$\vdots$$

$$\frac{\partial}{\partial a_N} \sum_x \sum_y \sum_z \| f(x,y,z,a_1,a_2,...a_N) - \begin{bmatrix} x \\ y \\ z \end{bmatrix} \|^2 = 0, \qquad (2c)$$



(a) data: (x_0, y_0, z_0), ..., (x_{20}, y_{20}, z_{20})

(b) model: $z = (x - 10)^2 + (y - 10)^2 + 5$

(c) sweet spot: $(x^*, y^*) = (10, 10)$

Figure 1: When data in (a) are modelable, a mathematical model in (b) is constructed to represent the data in (a), and an optimal point of the model can be calculated in (c).

where operators $\partial/\partial a_1$, $\partial/\partial a_2$, ..., $\partial/\partial a_N$ are partial derivative operators regarding the or s parameter to a_1, a_2, ..., a_N. Figure 1 (b) shows the result of modeling the data in 1 (a):

$z = (x - 10)^2 + (y - 10) 2 + 5$, (3)

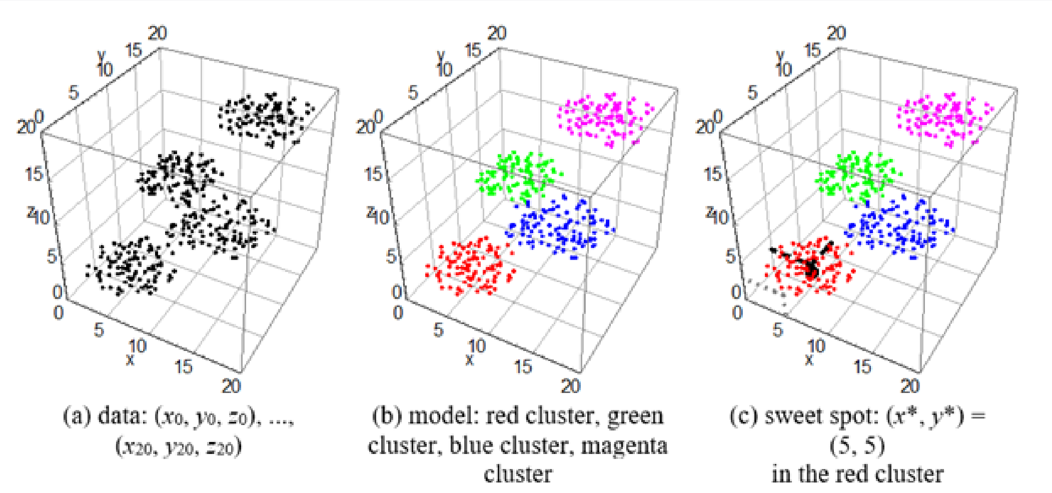
don d and the minimum value of the function z (x, and) are calculated routinely similar to the task of calculating the parameters a_1, a_2, ..., to a_N in (2) where the solution optimum is

$x * = 10, y * = 10, z \quad * = 5$, (4)

as shown in Figure 1 (c).

When the data are not formable s as shown in Figure 2 (a), look for another approach to determine the optimum point in these data. In this case, the use of data mining is recommended to discover the desired solution. In general, examining each data point can yield a point that contains the minimum value of the z component in the number sense, but this point could not represent the optimal point when there are many fluctuations in the data. For this reason, it is better to determine the sweet spot in a collective effort based on the data in a data cluster of a similar nature. Therefore, it is required to identify data clusters in which each cluster only contains data of a similar nature.

Figure 2 shows the case of non-modeling data with the use of the data mining approach to determine an optimal point. In this approach, clusters of data are identified in the segmentation step, with case results shown in Figure 2 (b). For each cluster identified in the segmentation step, its data is analyzed to find a representation of that cluster. In this rendering step, there are two methods: statistical method and deterministic method. Statistical method requires establishing a statistical model with a density distribution function where this function must be determined. The mean of the data is calculated according to its density distribution function. Deterministic method is simpler with the mean of the data being calculated according to the averaging function with an implication that the density of the data is uniformly distributed. Figure 2 (c) shows the cluster medium that represents the sweet spot.



(a) data: (x_0, y_0, z_0), ..., (x_{20}, y_{20}, z_{20})

(b) model: red cluster, green cluster, blue cluster, magenta cluster

(c) sweet spot: $(x^*, y^*) =$ (5, 5) in the red cluster

Figure 2: When data in (a) are not modelable, a cluster model in (b) is set up to represent the data in (a), and an optimal point of the model can be determined in (c) based on the analysis of each cluster.

Methodology

In this work, the data mining methodology is selected to determine the optimal point of a data set of the Granny Smith apple dehydration operation at the Agroindustrial Surfrut company, which is the sponsor of this project. Although the data mining methodology normally consists of four steps: segmentation, representation, compaction, and classification [12,13] In the context of determining the sweet spot, only the first two steps are required of segmentation (to determine clusters of data of a similar nature) and rendering (to determine the sweet spot in each of the clusters).

Due to the statistical nature of the data not known at the beginning of the project, the K-mean method [15] was selected, which is deterministic for data segmentation in clusters. This method has a computational advantage over the other deterministic method, hierarchical method [16], for segmentation. The K-medi method or is based on the distance formula to measure the similarity between two numeric data points. If this distance is small, these two numerical data points are said to be similar. When two data points are similar, they are assigned to the same cluster. Due to the use of the distance formula, the shape of each cluster tends to be a circle, a sphere, or a hypersphere [17,18].

The K-mean method begins with a set of data, and the assumption that there are N clusters in this set. Initially, the centers of these N clusters are assigned with arbitrary data (commonly with random numbers s). With this initial condition, the K-mean method is performed in iterations, with each iteration consisting of the following steps:

a) for each data point, the distance from this point to each center of the N clusters is calculated

b) the minimum distance is selected, and the point is assigned to the cluster that has the minimum distance to this data point

c) after all data points are assigned to the clusters, the center of each cluster is recalculated, and

d) if there are changes in the centers of the clusters, repeat steps (a) to (d)

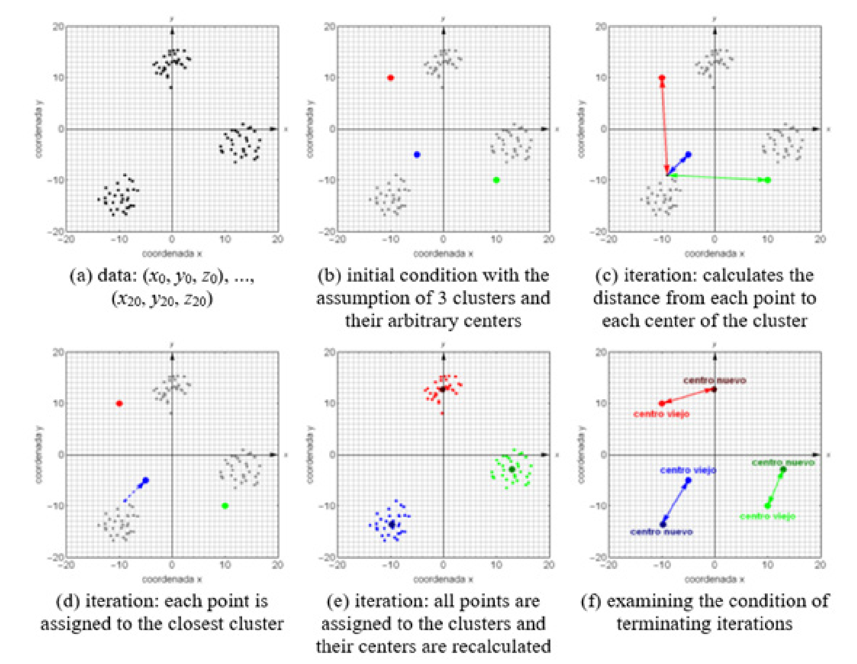
e) if there are no changes the iterations are terminated and the assignments of the points to the clusters are the final result

Figure 3 shows the algorithm of the K-mean method. In Figure 3 (a), a data set is considered for segmentation through the K-mean method. In Figure 3 (b), it is assumed that there are 3 clusters in the dataset, and the centers of these clusters start

How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643. DOI: 10.19080/TTSR.2020.04.555643

with arbitrary numbers (-10, 10) for the red cluster, (-5, -5) for the blue cluster, and (10, -10) for the green cluster. In Figure 3 (c), the data point (-9, -9) is considered with respect to the centers of the three clusters, with the distances from this point to the centers calculated. In Figure 3 (d), the distance from the point (-9, -9) to the center (-5, -5) of the blue cluster is selected because it is the minimum distance between the three distances calculated in Figure 3 (c) , and the point (-9, -9) is assigned to the blue cluster. In Figure 3 (e), all other data points are assigned to the clusters in a similar way, and then the centers of the clusters are recalculated. In Figure 3 (e), the termination condition is examined.

It is important to note that a guess of how many clusters exist within the dataset is required at the start of the K-mean method. In Figure 3 of 2-dimensional data, it is easy to verify the riddle with visual inspection. However, the purpose of using 2-dimensional data is to illustrate the concept of the K-mean method in which each step can be visualized and verified by visual inspection. In reality, when data is in many dimensions, it is not possible to verify through visual inspection whether the guess of how many clusters there are is correct. In this case, another verification technique must be developed. Considering the case of 2-dimensional data in Figure 4 (c). By visual inspection, it is clear that there are three clusters in the dataset. However, if the initial guess is wrong, there will be some overlapping clusters as shown in Figure 4 (d) and Figure 4 (f). When the initial guess is correct as in Figure 4 (e), there will be no overlapping clusters.



(a) data: (x_0, y_0, z_0), …, (x_{20}, y_{20}, z_{20})

(b) initial condition with the assumption of 3 clusters and their arbitrary centers

(c) iteration: calculates the distance from each point to each center of the cluster

(d) iteration: each point is assigned to the closest cluster

(e) iteration: all points are assigned to the clusters and their centers are recalculated

(f) examining the condition of terminating iterations

Figure 3: shows the algorithm of the K-mean method. In Figure 3 (a), a data set is considered for segmentation through the K-mean method. In Figure 3 (b), it is assumed.

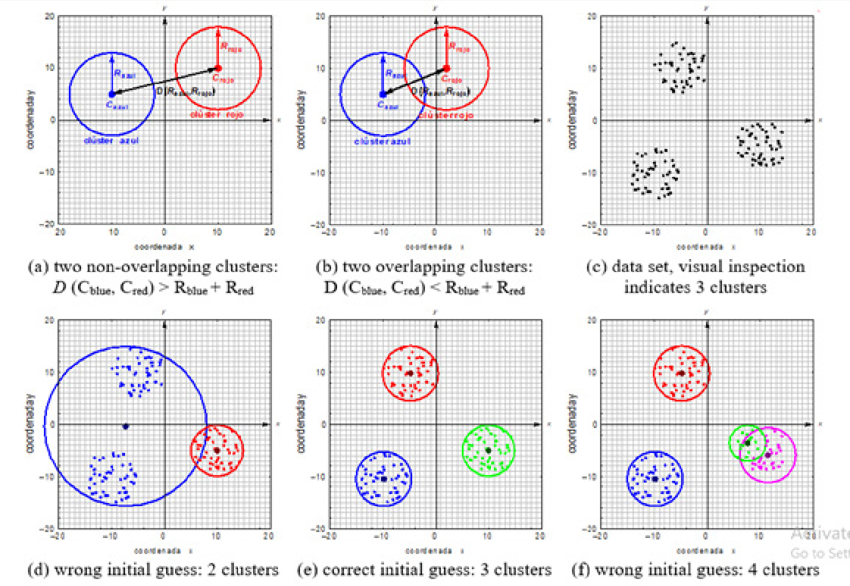
To detect whether the initial guess is correct or incorrect, the number of overlapping clusters must be considered in the results of segmenting a dataset. In Figures 4 (a) and 4 (b), these two overlapping and non-overlapping cluster scenarios are demonstrated. While these two scenarios are easy to verify with human eyes in cases of 2-dimensional or 3-dimensional data, it is much more difficult to do so in cases of 4 or more- dimensional data. For this reason, a simple method is developed to detect overlapping clusters through numerical analysis: for each cluster with its center and radius already determined, it is drawn a circle around said cluster as a boundary separating it from other clusters. In Figure 4 (a), when there is no overlap between two clusters, the distance between their centers is more than the sum of their radii. In Figure 4 (b), when there is overlap between two clusters, the distance between their centers is less than the sum of their radii. These descriptions are converted into mathematical expressions in the following:

Rule 1: $D(C_1, C_2) \geq R_1 + R_2 \Rightarrow$ there is no overlap, (5)

Rule 2: $D(C_1, C_2) < R_1 + R_2 \Rightarrow$ there is overlapping, (6)

Where $D(\cdot, \cdot)$ is the distance function between its two inputs, C_1 and C_2 are the centers of cluster 1 and cluster 2, and R_1 and R_2 are the radii of cluster 1 and cluster 2. Converting (5) and (6) in simple expressions that are implementable in computers, we obtain:

How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643. DOI: 10.19080/TTSR.2020.04.555643



(a) two non-overlapping clusters: $D\,(C_{blue}, C_{red}) > R_{blue} + R_{red}$

(b) two overlapping clusters: $D\,(C_{blue}, C_{red}) < R_{blue} + R_{red}$

(c) data set, visual inspection indicates 3 clusters

(d) wrong initial guess: 2 clusters

(e) correct initial guess: 3 clusters

(f) wrong initial guess: 4 clusters

Figure 4: When data in (a) are not modelable, a cluster model in (b) is set up to represent the data in (a), and an optimal point of the model can be determined in (c) based on the analysis of each cluster.

Rule 1a: $D\,(C_1, C_2) - R_1 - R_2 \geq 0 \Rightarrow$ there is no overlap, (7)

Rule 2b: $D\,(C_1, C_2) - R_1 - R_2 < 0 \Rightarrow$ there is overlap, (8)

The examples in Figures 4 (d), 4 (e), and 4 (f) can be analyzed by overlapping by examining the value of $\theta_{1,2} = D\,(C_1, C_2) - R_1 - R_2$. By this approach it is necessary to calculate the value $\theta_{n,m}$ for each pair of clusters n , m possible , and then identify the values of $\theta_{n,m}$, m that represent overlapping.

In Tables 1-3, the values of $\theta_{n,m}$ are calculated for three scenarios shown in Figures 4 (d), 4 (e), and 4 (4), respectively. In Table 1, there is overlap between cluster 1 and cluster 2. In Table 2, there is no overlap between the three clusters 1, 2, and 3. In Table 3, there is overlap between cluster 3 and cluster 4. Therefore, it is easy to conclude that the initial 3 cluster guess is correct because it results in no overlap between the resulting clusters. Considering the examples of the initial riddles to satisfy the requirement of the K-means method of knowing the cluster number, it is possible to arrive at the correct riddle by performing many segmentations, with each segmentation corresponding to a riddle and analyzing the values of $\theta_{n,m}$ of each segmentation to identify overlaps between the resulting clusters. The segmentation that contains no overlap or minimal overlaps is said to be the preferable segmentation because the initial guess is more likely to be correct than other segmentations.

Table 1: Value $\theta_{n,m}$ for Figure 4 (d).

	Cluster 1	Cluster 2
Cluster 1	--	**-2.35**
Cluster 2	**-2.35**	--

Cluster 1: C_1 = (-7.3, -0.3), R_1 = 15.3

Cluster 2: C_2 = (9.9, -5.0), R_2 = 4.9

Table 2: Value $\theta_{n,m}$ for Figure 4(e).

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	--	10.3	10. 4
Cluster 2	10.3	--	10. 7
Cluster 3	10. 4	10. 7	--

Cluster 1: C_1 = (-9.9, -10.4), R_1 = 5.2

Cluster 2: C_2 = (-4.8, 9.8), R_2= 5.3

Cluster 3: C_3 = (9.9, -5.0), R_1 = 4.9

Table 3: Value $\theta_{n,m}$ for Figure 4 (f).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	--	10.4	10.1	11.3
Cluster 2	10.4	--	9.5	12.0
Cluster 3	10.1	9.5	--	**-4.3**
Cluster 4	11.3	12.0	**-4.3**	--

Cluster 1: C_1 = (-9. 9, -10.4), R_1 = 5.2

Cluster 2: C_2 = (-4. 8, 9. 8), R_2 = 5.3

Cluster 3: C_3 = (7.6, -3.6), R_1 = 3.5

Cluster 4: C_4 = (11.4, -5. 9), R_1 = 5.3

Numerical Results

In this section, the actual data from the Agroindustrial Surfrut company are analyzed in the context of data mining described in the previous methodology section. This data has 13 fields for each record. Table 4 shows the names of the fields and their respective descriptions.

How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643. DOI: 10.19080/TTSR.2020.04.555643

Table 4: Fields of the Record in the Company Data Agroindustrial Surfrut.

Field Name	Course Description
X1	embedded code detailing the process
X2	describe the type of apple and cut it (in size)
X3	production start date
X4	date of production end
X5	amount of dehydrated product (kg.)
X6	amount of raw material processed (kg.)
X7	number of hours required for production
X8	yield of processed raw material (kg.)
X9	labor performance
X10	percentage of quantity
X 11	percentage shifted
X 12	operating mode standard
X 13	humidity level

Due to the requirement to know the number of clusters in the K-mean method of segmentation, this number has to be guessed. However, a divination is not true a correct divination. Therefore, segmentation is performed with many guesses, and segmentation is evaluated with minimal overlap between clusters results. In this work, the segmentation is done with five guesses of 2 clusters, 3 clusters, 4 clusters, 5 clusters, and 6 clusters. Overall, these numbers are commonly encountered in making segmentation: If a data set has more than 6 clusters, it is said that the whole is too fragmented and should not analyze it.

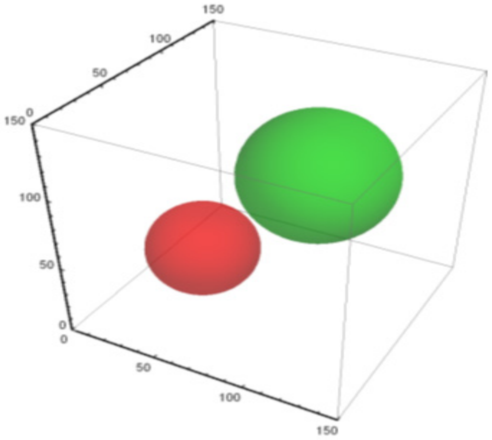
In the first case of 2 clusters, the results are shown in Table 5, with the cluster analysis shown in Table 6. Table 5 shows the centers of two clusters, where the values of the components X5, X6, ..., X12 represent the characteristic of a cluster. This characteristic is discussed later if the segmentation of the dataset into 2 clusters is evaluated to be the best representation. Table 6 shows the result of the normalized data on the scale from 0 to 100 for each component. In this table, the values of the centers of the clusters are shown in the normalized scale, and the analysis of overlaps is carried out in the calculation of the matrix of indicators $I_{n,m} = D (C_n , C_m) - R_n - R_m$, where $D (C_n , C_m)$ is the distance between the center of cluster C_n and the center of cluster C_m , R_n is the radius of cluster C_n , and R_m is the radius of cluster C_m . When $I_{n,m}$ (separation between cluster C_n and C_m) is negative, it means that there is an overlap between cluster C_n and cluster C_m . The segmentation with the least overlap is selected for the best guess on the number of clusters. Table 6 shows two clusters with order of magnitude spacing of 2 units. In this table, the separation of two clusters is (7-dimensional) is projected in the 3-dimensional environment for display.

Table 5: Cluster centers (based on non-normalized data) when guessing 2 clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	6675	71486	297.67	10.93	0.0466	10.835	0.0461
C2	1301	9204	68.31	7.22	0.0628	6.519	0.0464

Table 6: Centers two clusters and visualization of the projection 3D (based on normalized data.

	C1	C2	C3	C4
X5	44.09	8.60	na.	na.
X6	49.60	6.39	na.	na.
X7	41.00	9.41	na.	na.
X8	11.35	7.50	na.	na.
X9	3.71	5.00	na.	na.
X10	86.68	52.15	na.	na.
X12	36.33	36.54	na.	na.
Center Analysis (D (C_n, C_m) - R_n - R_m)				
	C1	C2	C3	C4
C1	0	2.72	na.	na.
C2	2.72	0.00	na.	na.
C3	na.	na.	na.	na.
C4	na.	na.	na.	na.



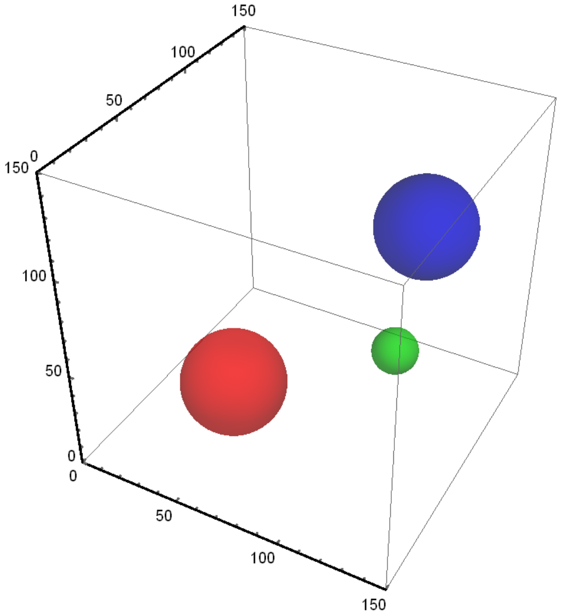
The s Tables 7 & 8 show results for segmentation-based divination 3 cluster. Again, the gaps between the clusters are projected into the 3-dimensional environment for display. These separations are in the order of magnitude of 20 to 75 units, much better than the case of 2 clusters. Tables 9 & 10 show results for segmentation based on guessing of 4 clusters. In this case, there is an overlap between cluster 1 and cluster 2, as indicated numerically in the matrix of gaps indicators, and in the graph of their projections in the 3-dimensional environment. Tables 11-13 show results for the 5-cluster guess-based segmentation. In case of 5 or more clusters, the separations cannot be projected in the 3-dimensional environment for display. However, the C5 cluster in this case has no data, meaning that there are only 4 possible clusters, and the result is the same as the result of the 4-cluster case, with overlap between cluster C1 and cluster C2. Similarly, in the case of 6 clusters in Tables 14-16, clusters C5 and C6 have no data, meaning that there are only 4 possible clusters, and the result is the same as the result of the case of 4 clusters, with overlapped between cluster C1 and cluster C2.

Table 7: Cluster centers (based on non-normalized actual data) when guessing 3 clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	1960	21145	102.80	11.456	0.0666	10.716	0.0543
C2	1192	2275	57.40	1.856	0.0515	1.508	0.0358
C3	7821	83426	342.71	10.822	0.0447	10.708	0.0436

Table 8: 3-cluster hubs and 3D projection visualization (based on normalized data).

	C1	C2	C3	C4
X5	12.95	7.8 8	51.66	
X6	14.67	1.58	57.88	
X7	14.16	7.91	47.20	
X8	11.90	1.93	11.24	
X9	5.31	4.11	3.57	
X10	85.73	12.07	85.67	
X12	42.82	28.24	34.34	
Center Analysis (D (C_n, C_m) - R_n - R_m)				
	C1	C2	C3	C4
C1	0.00	41.66	19.96	na.
C2	41.66	0.00	75.93	na.
C3	19.9 6	75.9 3	0.00	na.
C4	na.	na.	na.	na.



How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643.
DOI: 10.19080/TTSR.2020.04.555643

Table 9: Cluster centers (data real unstandardized) when guessing four clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	1746	19142	108.72	12.30 4	0.0776	11.274	0.0830
C2	2045	21452	98.42	10.529	0.0606	10.261	0.0412
C3	1224	2374	58.93	1.899	0.0516	1.168	0.0358
C4	7856	8391 5	10.84	10.84 4	0.0449	10.703	0.0433

Table 10: 4-cluster hubs and 3D projection visualization (normalized data).

	C1	C2	C3	C4
X5	11.53	13.51	8.09	51.89
X6	13.28	14.88	1.65	58.22
X7	14.97	13.56	8.12	47.50
X8	12.78	10.94	1.97	11.27
X9	6.18	4.83	4.11	3.58
X10	90.1 9	82.09	9.34	85.62
X12	65.36	32.49	28.20	34.12
Center Analysis ($D (C_n, C_m) - R_n - R_m$)				
	C1	C2	C3	C4
C1	0.00	**-8.28**	59.41	30.73
C2	**-8.28**	0.00	45.56	24.25
C3	59.41	45.56	0.00	80.33
C4	30.73	24.25	80.33	0.00

Table 11: Cluster centers (non-normalized actuals) when guessing 5 clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	1746	19142	108.72	12.304	0.0776	11.274	0.0830
C2	2045	21452	98.42	10.529	0.0606	10.261	0.0412
C3	1224	2374	58.93	1.899	0.0516	1.168	0.0358
C4	7856	83915	10.84	10.844	0.0449	10.703	0.0433
C5	0	0	0.00	0.000	0.0000	0.000	0.0000

Table 12: 5 cluster centers (normalized data).

	C1	C2	C3	C4	C5
X5	11.53	13.51	8.09	51.89	0.00
X6	13.28	14.88	1.65	58.22	0.00
X7	14.97	13.56	8.12	47.5	0.00
X8	12.78	10.94	1.97	11.27	0.00
X9	6.18	4.83	4.11	3.58	0.00
X10	90.19	82.09	9.34	85.62	0.00
X12	65.36	32.49	28.20	34.12	0.00

Table 13. Analysis of the 5 Centers (D (C_n, C_m) - R_n - R_m).

	C1	C2	C3	C4	C5
C1	0.00	**-8.28**	59.41	30.73	na.
C2	**-8.28**	0.00	45.56	24.25	na.
C3	59.41	45.56	0.00	80.33	na.
C4	30.73	24.25	80.33	0.00	na.
C5	na.	na.	na.	na.	0.00

Table 14: Cluster centers (based on non-normalized actual data) when guessing 6 clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	1746	19142	108.72	12.304	0.0776	11.274	0.0830
C2	2045	21452	98.42	10.529	0.0606	10.261	0.0412
C3	1224	2374	58.93	1.899	0.0516	1.168	0.035 8
C4	7856	83915	10.84	10.84 4	0.044 9	10.703	0.0433
C5	0	0	0.00	0.000	0.000	0.000	0.0000
C6	0	0	0.00	0.000	0.0000	0.000	0.0000

Table 15: 6 cluster centers (based on normalized data).

	C1	C2	C3	C4	C5	C6
X5	11.53	13.51	8.09	51.89	0.00	0.00
X6	13.28	14.88	1.65	58.22	0.00	0.00
X7	14.97	13.56	8.12	47.50	0.00	0.00
X8	12.78	10.94	1.97	11.27	0.00	0.00
X9	6.18	4.83	4.11	3.58	0.00	0.00
X10	90.19	82.09	9.34	85.62	0.00	0.00
X12	65.36	32.49	28.20	34.12	0.00	0.00

Table 16: Analysis of the 6 Centers (D (C_n, C_m) - R_n - R_m).

	C1	C2	C3	C4	C5	C6
C1	0.00	**-8.28**	59.41	30.73	na.	na.
C2	**-8.28**	0.00	45.56	24.25	na.	na.
C3	59.41	45.56	0.00	80.33	na.	na.
C4	30.73	24.25	80.33	0.00	na.	na.
C5	na.	na.	na.	na.	0.00	na.
C6	na.	na.	na.	na.	na.	0.00

Comparing the results between cases of 2 clusters, 3 clusters, 4 clusters, 5 clusters, and 6 clusters, the case of 3 clusters reaches the best separations between clusters (highest separations, and no overlaps). Therefore, it is concluded that the 3-cluster guess is appropriate, and the centers of the clusters can be analyzed in the context of identifying the cluster with the best operational performance.

Discussion

The centers of the three clusters extracted in the segmentation of the data set of the Agroindustrial Surfrut company on its operation of the apple dehydration process are summarized again in Table 17:

In this table, it is important to consider the variables X5 (amount in kg of dehydrated product), X6 (amount in kg of raw material processed), and X7 (number of hours required for production) to evaluate the operation. Because there are three variables under consideration, to determine the best cluster in terms of its operation, it is necessary to develop a formula that transforms these three variables into a scalar quantity for the purpose of comparing between these three clusters. To achieve efficiency, the number of hours per unit of processed fruit must be minimized, and the amount of water removed must be maximized. It is easy to calculate the number of hours per kg. of processed fruit by dividing the number of hours by the number of kilograms of processed fruit. Similarly, it is easy to calculate the amount of water removed by dividing the difference between the amount of fruit processed and the amount of dehydrated fruit by the amount of fruit processed, and the result is the percent mass of water removed. To combine the maximization of the amount of water removed and the minimization of the number of processing hours per kilogram of fruit, a linear combination with different weights is performed:

$$J = \alpha(\frac{X6 - X5}{X6}) - \beta(\frac{X7}{X6}) \qquad (1)$$

where α y β are weights representing the importance of these two criteria, y and J is the objective function to be optimized. Table 18 shows several scenarios with different sets of α y β to analyze the three clusters characterized by their centers shown in Table 17.

Table 17: Cluster centers (non-normalized actuals) with guessing of 3 clusters.

	X5	X6	X7	X8	X9	X10	X12
C1	1960	21145	102.80	11.456	0.0666	10.716	0.0543
C2	1192	2275	57.40	1.856	0.0515	1.508	0.0358
C3	7821	83426	342.71	10.822	0.0447	10.708	0.0436

Table 18: Calculation of the objective function J for various scenarios of weights α y β.

	X5	X6	X7	J (α = 0.3, β = 0.7)	J (α = 0.5, β = 0.5)	J (α = 0.7, β = 0.3)
C1	1960	21145	102.80	0.2687888	0.4512225	0.2697612
C2	1192	2275	57.40	0.1251516	0.2254066	0.1301978
C3	7821	83426	342.71	0.2690001	0.4510721	0.2698217

In Table 18, it is concluded that cluster C3 has the value of J optimized for α= 0.3 and β= 0.7, and also for α= 0.7 and β= 0.3. However, for α= 0.5 and β= 0.5, cluster C1 has the value of J optimized. Well, the selection of the weights α and β with emphasis on various criteria will determine the optimized solution. The products in optimized solution are related to their respective data in another data file with more operational information for the company to use to achieve its efficiency. This operating information will be used in the future to achieve similar efficiency.

Conclusion

In conclusion, data mining is effective in finding a solution to an optimization problem when there is an existing data set that represents the performance of a process, and this process cannot be modeled in a mathematical expression to formulate an optimization problem. in which its solution can be derived analytically. This approach is applied to the dehydration process of the fruit, with numerical results that will benefit the Agroindustrial Surfrut company that sponsors this project to optimize its operating cost.

References

1. Moore CA (2012) Automation in the Food Industry. New York, NY: Springer.
2. Torreggiani D (1993) Osmotic dehydration in fruit and vegetable processing. Food Research International 26(1): 59-68.
3. Abbott JA, Saftner RA, Gross KC, Vinyard BT, Janick J (2003) Consumer evaluation and quality measurement of fresh-cut slices of 'Fuji,' 'Golden Delicious,' GoldRush,' and 'Granny Smith' apples". Postharvest Biology and Technology 33(2): 127-140.
4. Shrouf F, Ordieres MJ, García SA, Ortega MM (2014) Optimizing the production scheduling of a single machine to minimize total energy consumption costs. Journal of Cleaner Production 67(15): 197-207.
5. Barney J (2014) Gaining and sustaining competitive advantage. Essex, UK: Pearson.
6. Emrouznejad A, Cabanda E (2014) Managing Service Productivity: Using Frontier Efficiency Methodologies and Multicriteria Decision Making for Improving Service Performance. New York, NY: Springer.
7. Alifah A (2017) Analysis of Liquidity, Leverage and Profitability in Assessing Financial Performance with Good Corporate Governance as Intervening Variables. Journal of Management 3(3).
8. Morgan P (2015) An Explanation of Constrained Optimization for Economists. Toronto, Canada: University of Toronto Press.
9. Birgin EG, Martínez JM (2014) Practical Augmented Lagrangian Methods for Constrained Optimization. Philadelphia, PA: Society for Industrial & Applied Mathematics.
10. Yong J (2018) Optimization Theory: A Concise Introduction. Hackensack, NJ: World Scientific Publishing Company.
11. Lasdon LS (2011) Optimization Theory for Large Systems. Mineola, NY: Dover Publications.
12. Aggarwal CC (2015) Data Mining: The Textbook. New York, NY: Springer.
13. Han J, Kamber M, Pei J (2011) Data Mining: Concepts and Techniques. Burlington, MA: Morgan Kaufmann.
14. Wang J, Wang J, Song J, Xu XS, Shen HT, et al. (2015) Optimized Cartesian K-means. IEEE Transactions on Knowledge and Data Engineering 27(1): 180-192.

How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643. DOI: 10.19080/TTSR.2020.04.555643

15. Wu J (2012) Advances in K-means Clustering: A Data Mining Thinking. Berlin, Germany: Springer- Verlag.

16. Devaraj S (2017) A Proposed New Algorithm for analysis of Hierarchical Clustering. Saarbrücken, Germany: Lap Lambert Academic Publishing.

17. Kirk DE (2004) Optimal Control Theory: An Introduction. Mineola, NY: Dover Publications.

18. Nise NS (2015) Control Systems Engineering. Hoboken, NJ: John Wiley & Sons.

This work is licensed under Creative Commons Attribution 4.0 Licens
DOI: 10.19080/TTSR.2020.04.555643

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats

(Pdf, E-pub, Full Text, Audio)

- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php

How to cite this article: Trung P, Teresita H T, Karla M C. Optimization of Raw Material Yield Using Data Mining. Trends Tech Sci Res. 2020; 4(4): 555643. DOI: 10.19080/TTSR.2020.04.555643