COMPARISON OF THE K-MEANS AND SELF-ORGANIZING MAPS TECHNIQUES TO LABEL AGRICULTURAL SUPPLY CHAIN DATA

Roberto F. Silva, Gustavo M. Mostaço, Fernando Xavier, Antonio Mauro Saraiva, Carlos E. Cugnasca

Department of Computer Engineering and Digital Systems,

Escola Politécnica da Universidade de São Paulo (USP), Brazil

roberto.fray.silva@gmail.com, gmostaco@usp.br, fxavier@usp.br, saraiva@usp.br, carlos.cugnasca@usp.br

ABSTRACT

The data produced in agricultural supply chains may be divided into three separated systems: (i) product identification and traceability, related to identifying production batches and places where the product has passed on the supply chain; (ii) environmental monitoring, considering mainly the temperature and relative humidity in storage and transportation; and (iii) processes, related to the data describing production processes and inputs used. Systems (i) and (ii) produce mainly structured data, while system (iii) produces non-structured data, and these are present in all agents in the supply chain. Data labeling on the different systems is an important step towards improving supply chain coordination and decision making related to traceability, production, and certification, among others. Nevertheless, it is a labor-intensive task, whose adoption is discouraged in data management activities. The main objective of this paper was to contribute to the reduction of interoperability problems by applying two clustering algorithms to label non-standardized data from agricultural supply chains. First, the data were clustered using k-means++ and self-organizing maps, with different model parameters. Then, a series of inferences were made to evaluate if the labels were correctly assigned, based on the characteristics of the data on each of the three systems. Lastly, a series of recommendations to improve the results of the models are discussed.

Keywords: agri-food, supply chains, unsupervised learning

1. INTRODUCTION

A supply chain (SC) can be described as a group of companies that are responsible for fulfilling the demands of a consumer segment (Chopra, Meindl, 2013). In the case of agricultural SCs, the products involved are agricultural products, such as grains, fruits, flowers, animal products, among others. An SC is divided into links or stages, which are groups of companies that are characterized by doing specific types of activities, such as farms, industries, logistics service providers, retailers, among others (Chopra, Meindl, 2013). Each company in an SC is termed an agent.

SCs generate a vast amount of heterogeneous data from different sources, processed on different systems and stored on different databases and formats (Corella, Rosale, Simarro, 2013). Generally, this data is not standardized among agents in the same link, and it can be processed and stored on different solutions, resulting in interoperability problems.

The data in an agricultural SC can be divided into three main systems: (i) product identification and traceability, related to identifying production batches and places where the product has passed among

the SC; (ii) environmental monitoring, considering mainly the temperature and relative humidity during the warehousing and transportation activities; and (iii) processes, related to the description of production processes and inputs used. Systems (i) and (ii) produce mainly structured data, while system (iii) produces mainly non-structured data (Pang et al., 2015; Verdouw et al., 2013; Chopra, Meindl, 2013; Verdouw et al., 2016).

As the Internet of Things (IoT) technologies become widely adopted in the SCs, the quantity of data that is generated and its associated problems tend to increase. Lack of interoperability, long periods of implementation and lack of communication are some of the problems of adopting this paradigm without proper standardization and coordination among agents (Harris, Wang, Wang, 2015).

Interoperability is a significant problem, as most SCs will probably continue to have coexisting technologies in terms of data generation, storage, and processing, resulting in data on different formats and resolutions. The complexity and time needed to standardize the data generated by all agents are prohibitive. Researches by Park and Song (2015) and Pang et al. (2015) consider the implementation of middlewares for the reduction of interoperability problems. Nevertheless, most of the research is theoretical and does not consider practical applications.

The main objective of this paper is to contribute to the reduction of interoperability problems by applying clustering algorithms to label non-standardized data from agricultural SCs. It is unpractical to manually label this data due to its volume and variety. Yet, this would lead to improvements in SC coordination and decision making related to traceability, production, and certification, among others. Our main research question is: "Can the K-means++ and self-organizing maps (SOM) models help on agricultural SC data labeling?" This is evaluated in light of the results from the models' implementations.

To the best of our knowledge, this is the first attempt to address this problem using unsupervised machine learning techniques. Two methods are implemented and studied: K-means++ and SOM. Logical inferences are developed to evaluate and improve the results. In addition, a framework for automatic data labeling will be briefly introduced.

The k-means is a clustering algorithm that was invented more than 50 years ago and is still used due to its good results (Jain, 2010). It basically creates points that will be used to calculate distances and form clusters, through a series of iterations. Nevertheless, it suffers from a problem due to the method used for its initialization, leading to local optima. To reduce the impact of this problem, the k-means++ algorithm was proposed (Arthur, Vassilvitskii, 2007). The k-means++ algorithm has a considerably lower processing time than using multiple random initializations.

The SOM algorithm is a neural network model for unsupervised machine learning and was created in 1982 (Kohonen, 1982). According to Kind and Brunner (2013), SOMs can be used to project n-dimensional data into a 2D map. It preserves the data topology, allowing the identification of hidden patterns that may not be observed in a k-means implementation, especially on datasets with a considerable amount of features. It can also be used for cluster analysis and labeling generation. Nevertheless, it is not nearly as used as the k-means algorithm.

2. METHODOLOGY

The methodology adopted for this work is divided into five main steps:

1. **Data collection and processing** using Python on the Jupyter IDE and LibreOffice Calc. It is composed of the following activities: **(i) data collection; (ii) data preprocessing** (unique identification of rows, separating target column from the data, dealing with missing data with zeros and interpolations, and .csv exporting); **(iii) data processing** (eliminate errors and outliers); **(iv) data fusion** (identify common features on the datasets, merge them, solve missing data problems and select which final features will enter the model); and **(v) data normalization** (using the MinMax scaler on Scikit-Learn package);

2. **Implementation of the K-means++ algorithm** using Scikit-Learn, considering seven implementations: K-means++, PCA with two features and five variations of k-means with random initialization (10, 20, 30, 40 and 50);
3. **Implementation of the SOM algorithm** using MiniSom[1], with the triangle neighborhood function and several combinations of the following parameters: learning rate (0.2, 0.3 and 0.5), sigma (3, 4, 5 and 6), and number of rounds (100, 500 and 1000);
4. **Definition and implementation of logical inferences** using Python in the Jupyter IDE, based on statistical analysis of the features and an in-depth literature review;
5. **Comparison of the implemented algorithms** using Scikit-Learn, divided into two parts: (I) unsupervised part, using silhouette score, homogeneity and processing time; and (ii) supervised part, using a confusion matrix and a classification report with the target labels.

Due to the lack of practical researches on clustering of agricultural SC data, and of suitable datasets that could be readily used in our research, some simplifying assumptions were made:

1. The predicted labels are related to the three main systems described in Section 1;
2. The considered SC is composed of four links: farm, industry, transportation, and retailer;
3. The dataset used for clustering was formed by the fusion of seven open datasets, described in Table 1. Each one represents partial data generated by different links in the SC as well as the most common information architecture present on it. This selection was based on a thorough search for open databases on Kaggle[2], Datahub[3] and Github[4]. Datasets features were analyzed during the data fusion activities;
4. The final dataset represents the situation of an agricultural SC, in which agents send their data to a common database in the cloud, without standardization.

Table 1. Open datasets selected to form the final dataset

Dataset number	SC Link and System	Description	Total number of features
01	Farm - 2	Environmental data collected[5]	9 (6 common, 3 unique)
02	Farm - 3	Herbicide application per plot[6]	10 (5 common, 5 unique)
03	Farm - 3	Plant growth measurements per plot[6]	12 (5 common, 7 unique)
04	Farm - 3	Productivity per plot[6]	6 (5 common, 1 unique)
05	Transportation - 1	Transportation from farm to industry[7]	9 (6 common, 3 unique)
06	Industry - 2	Environmental data at the industry[8]	15 (8 common, 7 unique)
07	Industry - 3	Operational data at the industry[9]	12 (5 common, 7 unique)

3. RESULTS AND DISCUSSION

Among the several implementations of k-means tested, the k-means++ and k-means random 50 times presented the best results. In relation to the SOMs implementations, the best results were achieved

1 https://github.com/JustGlowing/minisom
2 https://www.kaggle.com
3 https://datahub.io
4 https://github.com
5 INMET - Data from Rio Verde, GO - http://www.inmet.gov.br
6 Scribner et al (2003); Silva et al (2012); Omer et al (2015); Agridat - https://github.com/kwstat/agridat
7 Cashew Truck Arrivals - https://www.kaggle.com/extralime/cashew-truck-arrivals
8 Wsn-indfeat-dataset - https://github.com/apanouso/wsn-indfeat-dataset
9 Vega shrink-wrapper degradation - https://www.kaggle.com/inIT-OWL/vega-shrinkwrapper-runtofailure-data

with the following parameters: 3 (sigma), 0.2 (learning rate), and 500 (rounds). Table 2 contains the unsupervised quality metrics for the above mentioned methods.

From the analysis of these results, we can conclude that: (i) k-means++ was the best method due to a higher homogeneity score compared to the SOM model, it has a similar silhouette score compared to the k-means random 50 times, but presents significantly lower run time when compared to it; and (ii) the SOM model had a considerably worse silhouette score, and a much higher run time compared to k-means methods.

Some of the main reasons that may explain these results are: (i) the data imbalance among labels (System 1 contains 671 data points, while System 2 contains 20.527, and System 3 contains 15.153); (ii) the considerably small size of the dataset, especially for a neural network method; and (iii) the heterogeneity of data itself. Nevertheless, these results can work as a baseline for other implementations of unsupervised learning on the agricultural SCs domain.

Table 2. Model implementation results

Model	Processing time (s)	Homogeneity	Silhouette
K-means random 50 times	2.68	0.640	0.472
K-means++	0.41	0.640	0.426
SOM	52.50	0.569	0.012

After implementing the models and analyzing the unsupervised machine learning metrics, it was implemented a confusion matrix and a classification report, comparing the predicted clusters with the real labels. Table 3 contains the classification report for the k-means++ model and Table 4 contains the classification report for the SOM model. It is possible to observe that k-means++ can still be considered as the best model to cluster our dataset, presenting a considerably high precision, recall and F1-score across clusters 2 and 3. Nevertheless, it did not correctly classify cluster 1, missing all values. We believe that this is due to the data imbalance and the heterogeneity of this cluster. For the SOM model, although it manages to have a high recall for cluster 1, it achieves that by associating many more data points to this cluster than it really has (about 10 times more).

Table 3. Classification report for the k-means++ model

Quality metric	Cluster 1	Cluster 2	Cluster 3	Weighted average
Precision	0.00	0.86	1.00	0.90
Recall	0.00	1.00	0.81	0.90
F1-Score	0.00	0.93	0.90	0.90
Data points associated	250	23813	12288	-
Real number of data points	671	20527	15153	-

As a conclusion of both the unsupervised and supervised metrics analysis, the k-means++ model performed better than the SOM model. Nevertheless, this is an initial analysis of this domain, as stated before. More experiments with different datasets, are needed to validate this hypothesis. Based on the dataset that was created during our analyses, it is possible for other researchers to further improve on our findings.

Table 4. Classification report for the SOM model

Quality metric	Cluster 1	Cluster 2	Cluster 3	Weighted average
Precision	0.10	1.00	1.00	0.98

Recall	1.00	0.85	0.80	0.84
F1-Score	0.18	0.92	0.89	0.90
Data points associated	6589	18261	11501	-
Real number of data points	671	20527	15153	-

Based on our model implementation, we devised an initial framework for automatic data labeling for agricultural SCs. It is composed of the following steps:

1. **Data generation**, with standardization as the main target;
2. **Data collection**, with automatic row and dataset unique identification. This should, ideally, be done locally, if possible. However, as most agricultural SCs suffer from a lack of coordination, it could be executed automatically in the cloud;
3. **Data processing**, generating a new dataset without outliers, which are eliminated using logical inferences based on the features mean and standard deviation. Part of this process can be automatized. It is essential to maintain the original dataset, as some food quality problems are directly related to extreme variations on temperature, RH, gases, among others;
4. **Data fusion**, considering both the identification and selection of common features and the percentage of null cells in each feature. We believe this step can present some automation problems, as it would involve a considerable amount of inferences, which may not be clear before analyzing the datasets. More research is needed on this topic;
5. **Data normalization**, using the MinMax or standard scalers;
6. **Application of the labels on the original datasets**. This process can be automated, as the individual observations are uniquely identified. Future work is needed in order to identify if it is possible to extend the predicted labels to the observations that were considered initially as outliers and if any value is added to the SC by doing so.

4. Conclusions

In this paper, a dataset, representative of the data in an agricultural SC, was built and two models were implemented for clustering analysis to generate labels for these data: k-means++ and SOM. It was observed that the k-means++ performed better for both unsupervised and supervised machine learning metrics, while also having a considerably lower processing time. The results can be considered satisfactory, and a framework was proposed for automatic data labeling on this domain, becoming an important contribution of this work towards improving supply chain coordination and decision making in the agri-food sector.

Some limitations were observed, related to (i) the lack of open datasets that could be used to evaluate agricultural SCs, especially the ones containing data from the whole SC; and (ii) the lack of a framework to analyze heterogeneous, non-standardized data generated throughout agricultural SCs. Further work is needed to improve and enlarge the dataset. The implementation of other models, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Expectation-Maximization (EM) is also intended, as well as dealing with the current cluster imbalance in our data. The development of a common dataset, that can be used for evaluating data labeling in this domain, should be done in cooperation with partners from the different links of an agricultural SC.

Acknowledgments

This work was supported by Itaú Unibanco S.A. through the Itaú Scholarship Program, at the Centro de Ciência de Dados (C²D), Universidade de São Paulo, Brazil, by the National Council for Scientific and

Technological Development (CNPq), and also by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

REFERENCES

Arthur, D., Vassilvitskii, S. (2007) 'K-means++: The advantages of careful seeding', In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035.

Carrasco Kind, M., Brunner, R.J. (2014) 'SOMz: photometric redshift PDFs with self-organizing maps and random atlas', Monthly Notices of the Royal Astronomical Society, 438(4), pp.3409-3421.

Chopra, S., Meindl, P. (2013) 'Supply chain management: Strategy, planning, and operation', 5th ed. New Jersey, USA: Pearson Education, 528pp.

Corella, V.P., Rosalen, R.C., Simarro, D.M. (2013) 'SCIF-IRIS framework: a framework to facilitate interoperability in supply chains'. International Journal of Computer Integrated Manufacturing, 26(1-2), pp.67-86.

Harris, I., Wang, Y., Wang, H. (2015) 'ICT in multimodal transport and technological trends: unleashing potential for the future', International Journal of Production Economics, 159, pp.88-103.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', Pattern Recognition Letters, 31(8), pp.651-666.

Kohonen, T. (1982) 'Self-organized formation of topologically correct feature maps', Biological Cybernetics, 43(1), pp.59-69.

Omer, S.O., Abdalla, A.W.H., Mohammed, M.H., Singh, M. (2015) 'Bayesian estimation of genotype-by-environment interaction in sorghum variety trials'. Communications in Biometry and Crop Science, 10, pp.82-95.

Pang, Z., Chen, Q., Han, W., Zheng, L. (2015) 'Value-centric design of the internet-of-things solution for food supply chain: value creation, sensor portfolio and information fusion', Information Systems Frontiers, 17(2), pp.289-319.

Park, S., Song, J. (2015) 'Self-adaptive middleware framework for Internet of Things', In: 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE), pp. 81-82.

Scribner, E.A., Battaglin, W.A., Dietze, J.E., Thurman, E.M. (2003) 'Reconnaissance data for glyphosate, other selected herbicides, their degradation products, and antibiotics in 51 streams in nine Midwestern States', No. 2003-217.

Silva, A.M.D., Degrande, P.E., Suekane, R., Fernandes, M.G., Zeviani, W.M. (2012) 'Impacto de diferentes níveis de desfolha artificial nos estádios fenológicos do algodoeiro', Revista de Ciências Agrárias, 35(1), pp.163-172.

Verdouw, C.N., Vucic, N., Sundmaeker, H., Beulens, A. (2013) 'Future internet as a driver for virtualization, connectivity and intelligence of agri-food supply chain networks'. International Journal on Food System Dynamics, 4(4), pp.261-272.

Verdouw, C.N., Wolfert, J., Beulens, A.J.M., Rialland, A. (2016) 'Virtualization of food supply chains with the internet of things'. Journal of Food Engineering, 176, pp.128-136.