**OFFICIAL TEMPLATE FOR THE SUBMISSION OF PAPERS**
**II INTERNATIONAL SCIENTIFIC CONVENTION**
**"II ICC UCLV 2019"**

**JUNE 23th – 30th, 2019**
**CAYOS DE VILLA CLARA. CUBA.**

# WORKSHOP ON INTERNET OF THINGS WITH INTELLIGENT SYSTEMS AND DATA SCIENCES

## A Quality Measure for Multi-label Datasets on the Apache Spark Framework

**Ricardo Sánchez[1], Marilyn Bello[1,2], Carlos Morell[1], Rafael Bello[1], Koen Vanhoof[2]**

1-Computer Science Department, Universidad Central de Las Villas, Cuba,
risanchez@uclv.cu
2-Faculty of Business Economics, Hasselt University, Belgium

**Abstract:** In the last years, the amounts of data have increased considerably and therefore, it is becoming more complex to handle these volumes of information. Measuring the data quality is a pivotal aspect to assess the classifier's discriminatory power as the classifiers accuracy heavily depends on the data used to build the model. Multi-label classification is one specific type of classification problem, which has generated an increasing interest in recent years. However, there are no quality measures for multi-label datasets implemented in cluster computing frameworks to evaluate large datasets. This work aims to implement a measure of data quality for multi-label datasets based on Granular Computing under the Apache Spark framework. As a result, it was possible to calculate the values of the quality measure for the datasets, and even in relatively short times.

**Keywords:** Quality Measure, Multi-label Classification, Apache Spark.

### 1. Introduction

Multi-label classification is a particular case of classification (Tsoumakas, Katakis, & Vlahavas, 2010), where each example patterns has associated a vector of outputs (or labels), instead of only one value. The dimensionality treatment in multi-label datasets is

a more complex topic than in traditional classification (Herrera, Charte, Rivera, & Del Jesus, 2016). At present, one of the environments in which this type of data is frequently generated, characterized by a large amount of information, is the Internet of Things (IoT), in which there are a large number of data generation sources (Babbar & Schölkopf, 2017)(Gonzalez-Lopez, Cano, & Ventura, 2017)(Gonzalez-Lopez, Ventura, & Cano, 2018) (Moyano et al., 2019).

Apache Spark is a modern distributed computing framework, the basis of its architecture is a data structure which is a read-only multiset of data items distributed through a cluster of machines that is kept in a fault-tolerant environment. There are no algorithms or measures for the preprocessing stage of multi-label datasets in Apache Spark framework. Considering that the efficacy of the methods developed to solve multi-label classification problems depends on the quality of the data, it would be convenient to have a measure to do this. In this paper, we proposed a measure to overcome this issue on the framework of Apache Spark.

## 2. A Quality Measure for Multi-label Dataset and its Implementation on the Apache Spark Framework

From the perspective of the Granular Computing a relationship of interest is the relationship between the granulation of a universe according to the condition features and the granulation according to the decision classes. A measure that allows us to calculate this degree of similarity between two granulations is the Similarity Quality Measure (Cabrera, Pérez, Mota, & Jimenez, 2011). From this approach, new measures can be established. In Equation (1) we defined the *multi-label datasets quality measure*.

$$MLdQM = \frac{1}{|U|}\sum_{\forall x \in U} \frac{\sum_{\forall y \in U} 1 - |fsim(x,y) - dsim(x,y)|}{|U|} \quad (1)$$

where *fsim* and *dsim* represent the similarity degree between the objects *x* and *y* according to the condition features and decision classes respectively. The implementation as such

of the initial formulation results in an algorithm of high computational complexity. It is easily verifiable that the Equation (1) is equivalent to Equation (2):

$$MLdQM = 1 - \frac{1}{N^2}\sum_{x=0}^{N}\sum_{y=0}^{N} A(x,y)\,,\ A(x,y) = |fsim(x,y) - dsim(x,y)| \quad (2)$$

The implementation of the proposed data quality measure brings embedded the calculation of the similarity of all pairs of objects. We use the DIMSUM (Zadeh & Carlsson, 2014) algorithm for avoid computational complexities dependent on the dimension of the data. The sum of all the values in the matrix *A* is a reduction that is solved via partial sums that partially add data collaboratively, this operation is known as treeAggregate (Meng et al., 2016).

## 3. Results and discussion

The implemented algorithm was tested using three multi-label datasets taken from the MULAN repositories. For all datasets we used a Big Data cluster with Spark 2.3 in YARN mode as execution environment, using 10 nodes and 2 GB of RAM memory in each of these. The *MLdQM* values are obtained in relative short times. For example, in the case of the bibtex dataset to calculate the *MLdQM* value without a distributed computing approach takes approximately 48 hours on a computer with an Intel i5 (7th generation) and 16 GB of RAM while using the proposed algorithm takes 28 minutes.

## 4. Conclusions

In this work, the case of how to implement in the Spark environment the calculation of a measure of data quality for multi-label datasets developed from the perspective of Granular Computing is addressed. The proposed implementation is based on the programming primitives of Spark, and the experimentation shows how to work with multi-label datasets for which a local implementation is unaffordable.

For future work, we propose to develop alternatives to DIMSUM for other similarity measures.

## 5. References

Babbar, R., & Schölkopf, B. (2017). Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the tenth ACM international conference on web search and data mining* (pp. 721–729).

Cabrera, Y. F., Pérez, R. B., Mota, Y. C., & Jimenez, G. R. (2011). Improving the MLP learning by using a method to calculate the initial weights of the network based on the quality of similarity measure. In *Mexican International Conference on Artificial Intelligence* (pp. 351–362).

Gonzalez-Lopez, J., Cano, A., & Ventura, S. (2017). Large-scale multi-label ensemble learning on Spark. In *2017 IEEE Trustcom/BigDataSE/ICESS* (pp. 893–900).

Gonzalez-Lopez, J., Ventura, S., & Cano, A. (2018). Distributed nearest neighbor classification for large-scale multi-label data on spark. *Future Generation Computer Systems*, *87*, 66–82. https://doi.org/10.1016/J.FUTURE.2018.04.094

Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016). Multilabel classification. In *Multilabel Classification* (pp. 17–31). Springer.

Meng, X., Bradley, J., Street, S., Francisco, S., Sparks, E., Berkeley, U. C., … Hall, S. (2016). MLlib : Machine Learning in Apache Spark. *Journal of Machine Learning Research 17*, *17*, 1–7. https://doi.org/10.1145/2882903.2912565

Moyano, J., Gibaja, E., Ventura, S., Cano, A., Moyano, J. M., Gibaja, E. L., & Ventura, S. (2019). *Speeding Up Classifier Chains in Multi-label Classification Data Mining with More Flexible Representations View project New Problems in Knowledge Discovery: A Genetic Programming Approach View project Speeding Up Classifier Chains in Multi-label Classification*. Retrieved from https://www.researchgate.net/publication/331821964

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Data mining and knowledge discovery handbook. *Mining Multi-Label Data*.

Zadeh, R. B., & Carlsson, G. (2014). Dimension Independent Matrix Square using, 1–17.