Data quality measures based on granular computing for multi-label classification

Peer-reviewed author version

BELLO GARCIA, Marilyn; NAPOLES RUIZ, Gonzalo; VANHOOF, Koen & Bello, Rafael (2021) Data quality measures based on granular computing for multi-label classification. In: Information Sciences, 560 , p. 51 -67.

DOI: 10.1016/j.ins.2021.01.027
Handle: http://hdl.handle.net/1942/33621

Data Quality Measures based on Granular Computing for Multi-label Classification

Marilyn Bello[a,b,*], Gonzalo Nápoles[b,c], Koen Vanhoof[b], Rafael Bello[a]

[a] *Computer Science Department, Universidad Central de Las Villas, Cuba*
[b] *Faculty of Business Economics, Hasselt University, Belgium*
[c] *Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands.*

Abstract

Rough set theory is a granular computing formalism that allows analyzing a given dataset through well-defined measures. Some of these measures aim to characterize datasets used to discover knowledge, mostly in traditional classification problems. Measuring the data quality is pivotal to estimate beforehand the problem's difficulty since a classification model's accuracy heavily depends on the data quality. However, to the best of our knowledge, there are no measures devoted to analyzing the quality of multi-label datasets. In this paper, we propose six data quality measures for multi-label problems, which are based on different granular approaches. Some of these measures redefine the decision class concept, while others redefine the consistency concept. Moreover, we study the impact of the similarity threshold parameters and the distance functions on the behavior of these measures. The numerical simulations show a statistical correlation between the measures that redefine the consistency concept and the performance of the ML-kNN classifier.

Keywords: Multi-label Classification, Granular Computing, Rough Set Theory, Data Quality Measures

*Corresponding author
Email address: mbgarcia@uclv.cu (Marilyn Bello)

1. Introduction

Multi-label classification (MLC) refers to the case where an object is associated with more than one class simultaneously [1, 2, 3]. Let $mlDS = (U, A \cup L)$ be a multi-label decision system, where the set U is a non-empty finite set of objects, A is a non-empty finite set of attributes that describe each observation, and $L = \{l_1, l_2, \ldots, l_k\}$ is a non-empty finite set of labels such that the label domain is $L_i = \{0, 1\}$. The traditional classification task is generalized to the prediction of several labels simultaneously.

Multi-label learning is still in an early development stage with respect to other machine learning fields. Some measures have been defined to characterize multi-label datasets, such as label cardinality, label density, mean imbalance ratio, concurrence among labels, etc. [4, 5, 6]. Nevertheless, none of them are intended to measure the quality of the data.

The data quality analysis used for the knowledge discovery process is related to topics such as data complexity [7, 8] and metalearning. According to [9], the complexity of classification problems depends on factors such as class ambiguity, class overlap, and the complexity of class separation boundaries. On the other hand, metalearning is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning and data mining processes [10]. This definition emphasizes the notion of metaknowledge, that is, knowledge that relates the characteristics of datasets with the performance of the available algorithms.

Some research efforts [5, 11, 9] have been oriented to characterizing the data using granular computing principles while relating these characteristics to the performance of the classifiers. In [12] the author presented a measure —termed *quality of classification*— that is used to quantify the degree of consistency in a classification problem. Consistency is understood as the extent to which objects that are similar according to predictive attributes are associated with the same decision classes. This provides a measure to quantify the separability of the decision classes. The quality of classification measure is based on the Rough Set

Theory (RST) [13], which is probably the most suitable approach to deal with data inconsistency. In [14] the authors proposed an extension of this measure for decision systems in which the domain of prediction and condition variables can be both discrete and continuous. Caballero et al. [15] further studied the relationship between this measure and the performance of several classifiers, but the simulations were mainly dedicated to standard classification problems.

Within the MLC context, the literature reports some studies related to imbalance measurement [16] and other complexity facts, such as the concurrence among frequent and infrequent labels [4]. However, since the efficacy of MLC methods depends on the data used to build the model, it would be convenient to have some data quality measure shedding light on the prediction accuracy we can obtain. The lack of such a measure in the MLC literature served as the main driving force behind this research.

In this paper, we study the problem of estimating the consistency degree in multi-label data and propose measures to quantify the quality of data in MLC problems. Our proposal fills the gap that has existed until now to assess the quality in multi-label datasets since, as far as we know, existing data quality measures operate with traditional classification datasets. The first approach attempts to adapt the quality of classification measure to the multi-label scenario. This leads to three measures that differ in the way that the information granules are derived. While this sounds straightforward, a problem comes to light: a multi-label object can be associated with several labels simultaneously. We redefine the decision class concept to carry out the universe's granulation. The second approach is devoted to adjusting the consistency concept to the multi-bale setting. The numerical results using 12 multi-label datasets show that the proposed measures allow estimating the consistency of the MLC datasets. Moreover, their values are correlated with the prediction rates attained by the MLC algorithm adopted for the simulations.

The rest of the paper is organized as follows: Section 2 discusses important concepts to this research. Section 3 presents the new measures for assessing the quality of MLC dataset while Section 4 elaborates on the performance of the

proposed data quality measures. Section 5 formalizes some concluding remarks and future research directions to be explored.

2. Theoretical background

This section discusses important concepts to this study: the foundations of the classical and extended RST formalisms.

2.1. Rough Set Theory

The RST formalism allows handling uncertainty in the form of inconsistency. This theory involves two components: an information system and an inseparability relation. The former is defined as $IS = (U, A)$, where U denotes a non-empty finite set of objects, and A represents a non-empty finite set of attributes or features describing each object. The latter is defined as $R = \{(x, y) \in U \times U | \forall a \in A, x(a) = y(a)\}$ [17, 18].

Any subset $X \subseteq U$ can be approximated by using two crisp sets: the lower and the upper approximation [19]. They are defined as $A_*X = \{x \in U : [x]_A \subseteq X\}$ and $A^*X = \{x \in U : [x]_A \cap X \neq \emptyset\}$ respectively, where $[x]_A$ (i.e. an equivalence class) denotes the set of inseparable objects associated to x using an indiscernibility relation defined by A. The objects in A_*X are categorically members of X, while the objects in A^*X are possible members of the subset X. This model does not consider any tolerance of errors: if two inseparable objects belong to different classes, then the decision system will be inconsistent. The definition of indiscernibility as an equivalence relation is excessively strict, especially when it comes to numerical features.

2.2. Extended Rough Set Theory

The lack of flexibility of the classical RST becomes more serious for classification and decision-making problems having numerical attributes where the equivalence relation is less likely to hold. This issue can be solved by extending the concept of inseparability relation such that we can replace the equivalence

relation with a weaker binary relation [20]. Equation (1) shows an indiscernibility relation based on a similarity function,

$$R_1 : xRy \Longleftrightarrow \varphi(x, y) \geq \xi_1 \tag{1}$$

where $0 \leq \varphi(x, y) \leq 1$ is a similarity function. This weak binary relation states that objects x and y are deemed inseparable as long as their similarity degree $\varphi(x, y)$ exceeds a similarity threshold $0 \leq \xi_1 \leq 1$. This relation actually defines a similarity class $\bar{R}(x) = \{y \in U : yRx\}$ that replaces the equivalence class. In this paper, we assume that $\varphi(x, y) = 1 - \delta(x, y)$, where $\delta(x, y)$ denotes the distance between objects x and y.

Equation (2) and (3) show how to compute the lower and upper approximations, respectively, as described in [20],

$$A_*X = \{x \in U : \bar{R}(x) \subseteq X\} \tag{2}$$

$$A^*X = \bigcup_{x \in U : \bar{R}(x) \cap X \neq \emptyset} \bar{R}(x). \tag{3}$$

Another extension of the classical RST is the inclusion of the fuzzy approach to obtain more flexible models. Fuzzy-rough sets [21, 22] use a fuzzy similarity relation to replace the equivalence relation. The fuzzy relation [23, 24] quantifies the strength of the similarity relationship between two objects in the $[0, 1]$ interval. This implies that all objects in U are related to each other but with different membership degrees. The advantage of using a fuzzy relation is that the threshold in Equation (1) is not necessary.

3. Data quality measures for multi-label decision systems

In this section, we propose several measures to quantify the quality of multi-label datasets. The measures proposed in this paper allow estimating the dependence degree between the decision attributes (i.e., the set of labels) and the

predictive attributes by using different granular approaches. Using such estimation will allow us to quantify the consistency degree of the dataset. In that way, we could adjust our expectations about the performance of MLC algorithms on a particular dataset. In the end, we should not expect an algorithm to produce impressive prediction rates on poor-quality datasets. The data quality measures presented in this section differ from each other in the way they determine the similarity between predictive and decision granules.

3.1. Adaptations to the quality of classification measure

The first three measures presented next attempt to adapt the quality of classification measure [12] to the multi-label scenario. This RST-based data quality measure, defined in Equation (4), quantifies the percentage of objects that can be correctly classified in a decision system,

$$\gamma_A(Y) = \frac{\left|\bigcup_{i=1}^{k} A_* Y_i\right|}{|U|} \tag{4}$$

where Y represents the set of classes, Y_i the set of objects that belong to the i-th class, and k the number of classes of the problem. This measure produces values in the $[0, 1]$ interval such that one indicates that all indiscernible objects share the same decision classes (which indicates total consistency). In contrast, zero indicates a total inconsistency of the dataset.

Note that this measure was originally conceived for traditional pattern classification problems where each instance is associated with a single decision class. This implies that Y_i defines a partition over the universe.

Aiming at deriving new data quality measures based on the rationale of the quality of classification measure, we need to calculate the lower approximation associated with each decision class. However, this brings the problem of defining the concept of *decision class* in the multi-label context. In this paper, we consider the following variants:

- Each label combination is a decision value. For example, let $L = \{l_1, l_2, l_3\}$ denote the set of labels, then a combination of labels could be "101", point-

ing out that the object is associated with labels l_1 and l_3. Consequently, "101" defines a decision class, and all the objects that are associated with labels l_1 and l_3 belong to it. This approach is derived from the label powerset method [6], which transforms the multi-label problem into a single-label problem with a single class.

- Each label is considered a decision value such that all objects with that label belong to this decision class. According to this definition, in the example above, there would be three decision classes.
- Decision classes are derived using a clustering algorithm [25, 26]. In this case, we cluster the objects by considering only the labels such that each cluster would represent a decision class.

The first measure is based on the first variant, where each possible combination in the dataset is considered a decision class. Hence, we can easily compute the multi-classification of quality (MCQ) measure as the ratio between the objects that belong to the lower approximation with respect to the cardinality of the universe. Equation (5) formalizes this measure,

$$MCQ_A = \frac{\left|\bigcup_{i=1}^{k} A_* Y_i\right|}{|U|} \tag{5}$$

where k is the number of label combinations, and Y_i is the i-th decision class containing all objects associated with the i-th combination. This approach transforms the MLC problem into a traditional classification problem, and therefore, Equations (4) and (5) are equivalent.

The intuition of the second data quality measure is that each label represents a decision class. This means that the i-th decision class will contain all objects associated with the i-th label. For example, if the object x is associated with l_1, and x and y are inseparable, then y must also be associated with l_1. Otherwise, x does not belong to the lower approximation of l_1 because there is inconsistency.

Equation (6) defines this measure,

$$MCQ_B = \frac{\left|\bigcup_{i=1}^{k} A_* L_i\right|}{|U|} \tag{6}$$

where k denotes the number of labels, L_i is the i-th decision class that contains all objects that have the i-th label. Observe that the set of decision classes in a multi-label dataset generates a covering of the universe, not a partition. A covering is a family of non-empty finite subsets whose union is equal to the universe but their intersection may be non-empty.

The third measure is based on the idea that clusters generated from the labels are fair representatives of the decision classes. No particular clustering algorithm [25, 26] is needed for this measure. However, it requires a distance function for binary spaces [27, 28].

This measure is computed according to Equation (7), such that k represents the number of clusters, C_i is the i-th decision class that contains all the objects contained in the i-th cluster,

$$MCQ_C = \frac{\left|\bigcup_{i=1}^{k} A_* C_i\right|}{|U|}. \tag{7}$$

If the clustering is strict and requires the objects to be associated with the same labels, then the problem is reduced to the first variant, thus leading to a partition. In contrast, if the clustering is flexible, then this approach produces a covering of the universe of discourse.

3.2. Quality measure based on a granulation approach with thresholds

The measures introduced in the previous subsection attempt to adapt the quality of classification measure to the multi-label scenario. These measures establish a relation between the similarity classes of objects and the decision classes. This relationship could be generalized from other forms of granulation of the universe, such as the relationship between the granulation by condition and the granulation by decision.

The granulation process divides the universe of discourse into subsets of individual objects (granules) that share similar properties [17]. In this process, informative relationships emerge when relating the granules formed from predictive features (predictive granules) with the ones derived from decision features (decision granules). From the perspective of data consistency, it is reasonable to assume that predictive and decision granules should be related to each other, i.e., there is some similarity between them [29]. Thus, we could define data quality measures for multi-label datasets by measuring the similarity between the granulation by condition and decision.

The rationale of the measure proposed in this subsection consists of measuring the extent to which the granules by condition and decision are similar. The granule by condition of an object $x \in U$ consists of all objects inseparable to x when considering the condition attributes. In contrast, the granule by decision can be defined as the set of inseparable objects to x when considering the labels. For each object in the dataset, the similarity degree between both granules can be calculated using the following equation,

$$\alpha_B(x) = \frac{\mid COND(x) \cap DEC(x) \mid}{0.5 \mid COND(x) \mid + 0.5 \mid DEC(x) \mid} \tag{8}$$

where $COND(\cdot)$ is the granule by condition and $DEC(\cdot)$ is the granule by decision. These granules can be defined by using the indiscernibility relation defined in Equations (1) and (9), respectively,

$$R_2 : xRy \iff \vartheta(x, y) \geq \xi_2 \tag{9}$$

such that $0 \leq \vartheta(x, y) \leq 1$ is a similarity function, and $\vartheta(x, y) = 1 - H(x, y)$, where H is the normalized Hamming distance (see Definition 1) between the label sets associated to the object x and y, respectively [30].

Definition 1. *Given two vectors $x, y \in \mathbb{R}^n$ we define the Hamming Distance between x and y, $H(x, y)$, to be the number of places where x and y differ.*

The relation R_2 states that x and y are deemed inseparable as long as their

similarity degree $\vartheta(x, y)$ exceeds a similarity threshold $0 \leq \xi_2 \leq 1$. After computing the similarity between the granules $COND(\cdot)$ and $DEC(\cdot)$, we can calculate the MCQ measure as follows:

$$MCQ_D = \frac{\sum_{\forall x \in U} \alpha_B(x)}{|U|}. \tag{10}$$

Larger measure values suggest consistency in the dataset. This means that the granules by condition and decision lead to similar coverings of the universe of discourse. At the same time, having larger consistency values is a reasonable heuristic indicator for the algorithms to perform well.

3.3. Quality measures based on a granulation approach without thresholds

The reader can observe that the granules $COND(\cdot)$ and $DEC(\cdot)$ are built using similarity relations involving a similarity function and the similarity threshold parameters. These components might have a significant impact on the granulation process, thus leading to quite dissimilar results. To suppress the need for the threshold parameters, we could only use the degree to which an object is similar to the others according to some ordinal scale.

An alternative to do this is using *rankings*. A ranking establishes the order of a set of objects based on a value associated with them. Hence, two rankings (i.e., by condition $R_c(x)$ and by decision $R_l(x)$) could be established for each object x based on its degrees of similarity with respect to the others according to the condition attributes (c) and the labels (l), respectively. These rankings can be compared using a distance function.

In this approach, the rankings $R_c(x)$ and $R_l(x)$ contain all objects ordered according to their similarity values with respect to the object x. The MCQ measure in Equation (11) formalizes this idea,

$$MCQ_E = \frac{\sum_{\forall x \in U} 1 - d(R_c(x), R_l(x))}{|U|} \tag{11}$$

where $d(.,.)$ is the normalized Spearman distance [31], which is given as follows,

$$d(\sigma, \tau) = \sum_{i \in U} |\sigma(i) - \tau(i)| \tag{12}$$

such that σ and τ denote the rankings generated from a finite set of objects, and $\sigma(i)$ and $\tau(i)$ represent the position (or order) of the i-th object in the rankings σ and τ, respectively. Note that several values with the same position might be observed in R_c or R_l. This must be considered when implementing the distance between the rankings to ensure consistent results.

Let us suppose the ranking generated by an object with respect to the condition features is the same as the one for the labels. In that case, we could infer that the relation between the objects according to the predictive features and the labels is similar. Thus, we can say that the data is consistent with a specific degree, which is given by the distance between both rankings.

Another alternative to avoid using the similarity thresholds when granulating the information space is to replace the hard similarity relations with fuzzy ones. As a result, the granules by condition $COND(\cdot)$ and decision $DEC(\cdot)$ will have soft boundaries (fuzzy sets to which all objects belong to with some degree). This approach is similar to the fuzzy-rough sets [22, 32] since the fuzzy relation replaces the crisp similarity function. In a nutshell, this alternative requires to rewrite the crisp similarity relations R_1 and R_2 as fuzzy relations as depicted in Equations (13) and (14), respectively,

$$xR_1y \Longleftrightarrow \varphi(x, y) \tag{13}$$

$$xR_2y \Longleftrightarrow \vartheta(x, y). \tag{14}$$

The fuzzy relations R_1 and R_2 are computed using the condition attributes and the similarity function φ, and the decision attributes and the similarity function ϑ, respectively. Therefore, for any object in the universe of discourse, two fuzzy sets $N_1(x)$ and $N_2(x)$ are built (the former is based on the condition

features and the latter is based on the decision features). These fuzzy sets are defined in Equations (15) and (16), respectively,

$$N_1(x) = \{(y, \mu_{R_1}(x, y)) \, \forall y \in U\} \tag{15}$$

$$N_2(x) = \{(y, \mu_{R_2}(x, y)) \, \forall y \in U\} \tag{16}$$

where $\mu_{R_1}(x, y)$ and $\mu_{R_2}(x, y)$ denote the membership degrees of the object y to $N_1(x)$ and $N_2(x)$, respectively, with $\mu_{R_1}(x, y) = \varphi(x, y)$ and $\mu_{R_2}(x, y) = \vartheta(x, y)$. Finally, Equation (17) shows how to compute the MCQ measure according to the fuzzy sets $N_1(x)$ and $N_2(x)$,

$$MCQ_F = \frac{1}{|U|} \sum_{\forall x \in U} \frac{\sum_{\forall y \in U} 1 - |\mu_{R_1}(x, y) - \mu_{R_2}(x, y)|}{|U|}. \tag{17}$$

Overall, the intuition of this measure is that rather consistent multi-label problems should generate similar fuzzy sets in the granulation process. This information can be understood as a data quality measure that does not depend on the similarity threshold parameters.

3.4. A simple example

In this subsection, we present a toy example illustrating how to compute the proposed measures. Let us suppose we have a multi-label decision system (as depicted in Table 1) with four objects $\{x, y, z, w\}$, three attributes $\{a_1, a_2, a_3\}$ and three labels $\{l_1, l_2, l_3\}$ such that the label sets $\{l_1, l_3\}$ and $\{l_1, l_2\}$ overlap. Table 2 portrays the similarity values between those objects according to the attributes and labels derived from the similarity functions φ and ϑ, respectively. Moreover, the clusters obtained with k-means algorithm are $C_1 = \{x, z, w\}$ and $C_2 = \{y\}$. In this example, we used the Hamming distance for handling objects in the label space [27]. Observe that objects y and w are inseparable but they are associated with different label sets.

MCQ_A reports a value of 0.5. The related knowledge structures are given as follow: the decision classes are $Y_{101} = \{x, z, w\}, Y_{110} = \{y\}$, the similarity classes

Table 1: Example of a multi-label dataset.

	$a1$	$a2$	$a3$	$l1$	$l2$	$l3$
x	5	1	648	1	0	1
y	3	1	29	1	1	0
z	6	3	0	1	0	1
w	3	1	30	1	0	1

Table 2: Similarity values between object by condition (left) and decision (right).

	x	y	z	w
x	1	0.2	0	0.2
y	0.2	1	0.18	0.99
z	0	0.18	1	0.18
w	0.2	0.99	0.18	1

	x	y	z	w
x	1	0.33	1	1
y	0.33	1	0.33	0.33
z	1	0.33	1	1
w	1	0.33	1	1

are $\bar{R}(x) = \{x\}, \bar{R}(y) = \{y, w\}, \bar{R}(z) = \{z\}, \bar{R}(w) = \{w, y\}$ with $\xi_1 = 0.75$, and the lower approximations are $B_*Y_{101} = \{x, z\}$, $B_*Y_{110} = \{\}$.

MCQ_B reports a value of 1. The related knowledge structures are given as follow: the decision classes are $Y_{l1} = \{x, y, z, w\}, Y_{l2} = \{y\}, Y_{l3} = \{x, z, w\}$, the similarity classes are $\bar{R}(x) = \{x\}, \bar{R}(y) = \{y, w\}$, $\bar{R}(z) = \{z\}, \bar{R}(w) = \{w, y\}$ with $\xi_1 = 0.75$, and the lower approximations are $B_*Y_{l1} = \{x, y, z, w\}$, $B_*Y_{l2} = \{\}, B_*Y_{l3} = \{x, z\}$.

MCQ_C reports a value of 0.5. The related knowledge structures are given as follow: the decision classes are $Y_{C_1} = \{x, z, w\}, Y_{C_2} = \{y\}$, the similarity classes are $\bar{R}(x) = \{x\}, \bar{R}(y) = \{y, w\}, \bar{R}(z) = \{z\}$, $\bar{R}(w) = \{w, y\}$ with $\xi_1 = 0.75$, and the lower approximations are $B_*Y_{C_1} = \{x, z\}, B_*Y_{C_2} = \{\}$.

MCQ_D reports a value of 0.52. The related knowledge structures are given as follow: the granules by condition are $COND(x) = \{x\}, COND(y) = \{y, w\}$, $COND(z) = \{z\}, COND(w) = \{w, y\}$ where $\xi_1 = 0.75$, and the granules by decision are $DEC(x) = \{x, z, w\}, DEC(y) = \{y\}$, $DEC(z) =$

$\{z, x, w\}, DEC(w) = \{w, x, z\}$ where $\xi_2 = 0.5$.

MCQ_E reports a value of 0.44. The related knowledge structures are given as follow: the ranking by condition are $R_c(x) = \{x \prec w \prec y \prec z\}, R_c(y) = \{y \prec w \prec x \prec z\}, R_c(z) = \{z \prec y \prec w \prec x\}, R_c(w) = \{w \prec y \prec x \prec z\}$, and the ranking by decision are $R_l(x) = \{(x \wedge z \wedge w) \prec y\}, R_l(y) = \{y \prec (x \wedge z \wedge w)\}, R_l(z) = \{(x \wedge z \wedge w) \prec y\}, R_l(w) = \{(x \wedge z \wedge w) \prec y\}$ where $O_i \prec O_j$ means that O_i proceeds O_j in the ranking and $(O_i \wedge O_j)$ indicates an equal score in the ranking for O_i and O_j.

MCQ_F reports a value of 0.56. The related knowledge structures are given as follow: the fuzzy sets by condition are $N_1(x) = \{(x, 1), (y, 0.2), (z, 0), (w, 0.2)\}$, $N_1(y) = \{(x, 0.2), (y, 1), (z, 0.18), (w, 0.99)\}, N_1(z) = \{(x, 0), (y, 0.18), (z, 1), (w, 0.18)\}, N_1(w) = \{(x, 0.2), (y, 0.99), (z, 0.18), (w, 1)\}$, and the fuzzy sets by decision are $N_2(x) = \{(x, 1), (y, 0.33), (z, 1), (w, 1)\}, N_2(y) = \{(x, 0.33), (y, 1), (z, 0.33), (w, 0.33)\}, N_2(z) = \{(x, 1), (y, 0.33), (z, 1), (w, 1)\}$, $N_2(w) = \{(x, 1), (y, 0.33), (z, 1), (w, 1)\}$.

3.5. Further discussion

As mentioned, the proposed data quality measures allow quantifying the quality of a multi-label decision system by assessing its consistency. In a multi-label context, consistency can be understood as the relationship between objects according to their predictive features and labels. The difference from one measure to another resides in the way this relationship is determined. The main features of each measure are highlighted below:

- MCQ_A is based on the quality of classification measure. Each combination of labels is considered a decision class.

  Advantages:

  – It allows using the quality of classification measure in the MLC setting without any modification.

  Disadvantages:

  – It requires to set up a threshold to compute the granules.

- The value of this measure depends on the total inclusion of the conditional granules into the decision granules.
- The computational complexity is $O(k|A||U|^2)$, where $|U|$ is the cardinality of the universe, $|A|$ is the cardinality of attribute set, and k the number of label combinations.

- MCQ_B is based on the quality of classification measure. Each label is considered a decision class.

  Advantages:

  - It allows using the quality of classification measure in the MLC setting without any modification.

  Disadvantages:

  - It requires to set up a threshold to compute the granules.
  - The value of this measure depends on the total inclusion of the conditional granules into the decision granules.
  - It is not considered the possible correlation among the labels.
  - The computational complexity is $O(k|A||U|^2)$, where $|U|$ is the cardinality of the universe, $|A|$ is the cardinality of the attribute set, and k is the number of labels.

- MCQ_C is based on the quality of classification measure. Each decision class is a cluster of similar objects according to the label space.

  Advantages:

  - It allows using the quality of classification measure in the MLC setting without any modification.

  Disadvantages:

  - It requires to set up a threshold to compute the granules.
  - The value of this measure depends on the total inclusion of the conditional granules into the decision granules.
  - It requires a clustering method to define the decision classes.
  - The computational complexity is $O(k|A||U|^2)$, where $|U|$ is the cardinality of the universe, $|A|$ denotes the cardinality of attribute set, and k the number of clusters.

- MCQ_D is based on the similarity relation between the granules by condition and by decision. Its value depends on the extent to which the granules by condition and by decision match.

  Advantages:

  - It is not necessary to define the decision classes.

  Disadvantages:

  - It is necessary to establish two thresholds to calculate the granulation by condition and by decision.
  - The computational complexity is $O(|U|^3 max\{|A|, |L|\})$, where $|U|$ is the cardinality of the universe, $|A|$ and $|L|$ are the cardinality of the attribute set and label set, respectively.

- MCQ_E is based on the similarity relation between the granules by condition and by decision. Its value depends on the similarity between rankings by condition and decision of each object.

  Advantages:

  - It is not necessary to define the decision classes.
  - It is not necessary to define any similarity threshold when computing the information granules.

  Disadvantages:

  - The construction of rankings by condition and decision could be computationally demanding.
  - The computational complexity is $O(|U|^2 max\{|A|, log(|U|), |L|\})$, where $|U|$ is the cardinality of the universe, $|A|$ and $|L|$ are the cardinality of the attribute set and label set, respectively.

- MCQ_F is based on the similarity relation between the granules by condition and by decision. Its value depends on the similarity between the fuzzy sets by condition and decision of each object.

  Advantages:

  - It is not necessary to define the decision classes.
  - It is not necessary to define any similarity threshold when computing the information granules.

Disadvantages: The computational complexity is $O(|U|^2 max\{|A|, |L|\})$, where $|U|$ is the cardinality of the universe, $|A|$ and $|L|$ are the cardinality of the attribute set and label set, respectively.

4. Numerical experiments and discussion

The purpose of this section is to study the behavior of the proposed measures in different multi-label datasets. Firstly, we compute the values of the measures for different parameter settings to evaluate the impact of these parameters on the results. Afterward, we analyze the correlation between those measure values and the performance achieved by the ML-kNN classifier [33, 34]. This empirical analysis allows us to conclude that three out of six measures correlate with algorithms' performance in MLC settings.

4.1. Dataset characterization

We leaned upon 12 multi-label datasets corresponding to three application areas in which multi-label data is frequently observed: text categorization, multimedia classification and bioinformatics. All datasets were taken from the MULAN [35] and RUMDR [36] repositories.

Table 3 outlines the number of objects ($|U|$), nominal attributes (*nominal*), numerical attributes (*numeric*), and labels for each dataset ($|L|$). The number of distinct label sets ($LSet$), calculated as the number of distinct combinations of labels found in the dataset is also given.

The TCS metric is often used as a theoretical complexity indicator [5]. This measure is calculated as the product of the number of attributes, the number of labels, and the number of distinct label sets. In order to avoid working with very large values, whose interpretation and comparison would be difficult, the *log()* function is used to adjust the scale resulting from the previous product. Hence, the larger the value, the more complex the processing of the dataset [5]. Remark that TCS values are logarithmic, so a difference of only one unit implies one order of magnitude smaller or larger.

Table 3: Characterization of the multi-label datasets.

Dataset	Domain	$\|U\|$	numeric	nominal	$\|L\|$	Lset	TCS
bibtex	text	7,395	0	1,836	159	2,856	20.54
birds	audio	645	258	2	19	133	13.39
cal500	music	502	68	0	174	502	15.59
corel5k	images	5,000	0	499	374	3,175	20.20
emotions	music	593	72	0	6	27	9.36
enron	text	1,702	0	1,001	53	753	17.50
flags	images	194	10	9	7	54	8.87
genbase	biology	662	0	1,186	27	32	13.84
medical	text	978	0	1,449	45	94	15.62
scene	images	2,407	294	0	6	15	10.18
slashdot	text	3,785	1,079	0	22	156	15.12
yeast	biology	2,417	103	0	14	198	12.56

4.2. Assessing classifier performance with Hamming Loss

The literature reports several measures to quantify the performance of MLC models such that accuracy, precision, recall, F-measure, among others. However, the *Hamming Loss* (HL) formalized in Equation (18) is probably the most used performance metric [1, 37] in MLC settings,

$$HL = \frac{1}{|U|}\frac{1}{|L|}\sum_{i=1}^{|U|}|Y_i \Delta Z_i| \tag{18}$$

where the Δ operator returns the symmetric difference between Y_i (the real labelset of the i-th object) and Z_i (the predicted one). Observe that, since the mistakes counter is divided by the number of labels, this metric will result in different assessments for the same amount of errors when used with datasets having a label set with a large cardinality.

4.3. Heterogeneous distance functions

Neighborhood measures characterize the superposition of classes by analyzing the local vicinity of the data points. In that regard, the similarity function

plays a key role. In [38] the authors studied several distance functions (which are the complement of the similarity functions). Such functions allow comparing the dissimilarity between two heterogeneous objects, i.e., objects comprising both numerical and nominal attributes.

Let $A = \{a_1, \ldots, a_M\}$ denote the attribute set, where a_j can be either numerical or nominal, and it has a weight $0 \leq \omega_j \leq 1$ quantifying its relevance. The similarity function $\varphi(x, y)$ between two objects x and y can be computed using one of the following distance functions:

- The Heterogeneous Euclidean-Overlap Metric (HEOM) in Equations (19) and (20) computes the normalized Euclidean distance between numerical attributes and an overlap metric for nominal attributes,

$$\delta(x, y) = \sqrt{\frac{\sum_{j=1}^{|A|} \omega_j \sigma_j(x, y)}{\sum_{j=1}^{|A|} \omega_j}} \tag{19}$$

where

$$\sigma_j(x, y) = \begin{cases} 0 & \text{if } a_j \text{ is nominal } \wedge\, x(j) = y(j) \\ 1 & \text{if } a_j \text{ is nominal } \wedge\, x(j) \neq y(j) \\ (x(j) - y(j))^2 & \text{if } a_j \text{ is numerical.} \end{cases} \tag{20}$$

- The Heterogeneous Manhattan-Overlap Metric (HMOM) is similar to the HEOM function, but it replaces the Euclidean distance with the Manhattan distance when computing the dissimilarity between two numerical values. Equations (21) and (22) show this distance function,

$$\delta(x, y) = \frac{\sum_{j=1}^{|A|} \omega_j \sigma_j(x, y)}{\sum_{j=1}^{|A|} \omega_j} \tag{21}$$

where

$$\sigma_j(x, y) = \begin{cases} 0 & \text{if } a_j \text{ is nominal } \wedge\, x(j) = y(j) \\ 1 & \text{if } a_j \text{ is nominal } \wedge\, x(j) \neq y(j) \\ |x(j) - y(j)| & \text{if } a_j \text{ is numerical.} \end{cases} \tag{22}$$

The reader can notice that the main difference between these distance functions is that the latter replaces the squared difference with the absolute difference when computing the dissimilarity between numerical attributes. Hence, it is reasonable to expect the HEOM distance to produce smaller values than those produced by the HMOM distance.

4.4. Computing the consistency value

Aiming at visualizing the simulation results, we split the proposed data quality measures into two groups. The first one includes measures MCQ_A, MCQ_B and MCQ_C, while the second group contains measures MCQ_D, MCQ_E and MCQ_F. The criteria for forming these groups are derived from the semantics of the proposed measures discussed in Section 3.

Figures 1 and 2 show the consistency values attached to the first group of measures for the HEOM and HMOM distance functions, respectively. Similarly, Figures 3 and 4 display the consistency values for the second group of measures. In our simulations, we arbitrarily set the similarity thresholds to $\xi_1 = 0.75$ and $\xi_2 = 0.5$, although other values are possible. Later on, we will study the effect of these parameters on the proposed measures.

Figures 1 and 2 show that the consistency values reached by the measures MCQ_A, MCQ_B and MCQ_C are small in most datasets. Usually, larger values of these measures are achieved for datasets having smaller *TCS* values (such as *bird*, *flag* and *emotion*). The values obtained when using the HEOM function are larger than the ones when using the HMOM function.

Figures 3 and 4 indicate that the consistency values obtained by MCQ_D, MCQ_E and MCQ_F are larger when compared with the ones produced by the measures in the first group. The simulation results suggest that the more complex the problem (with respect to the *TCS* values reported in Table 3), the larger the consistency value. This behavior holds for *bibtex*, *corel5k*, *medical* and *slashdot*. In the first group, the opposite happens. For this second group of measures, the values obtained when using the HMOM function are larger than the ones when using the HEOM function.

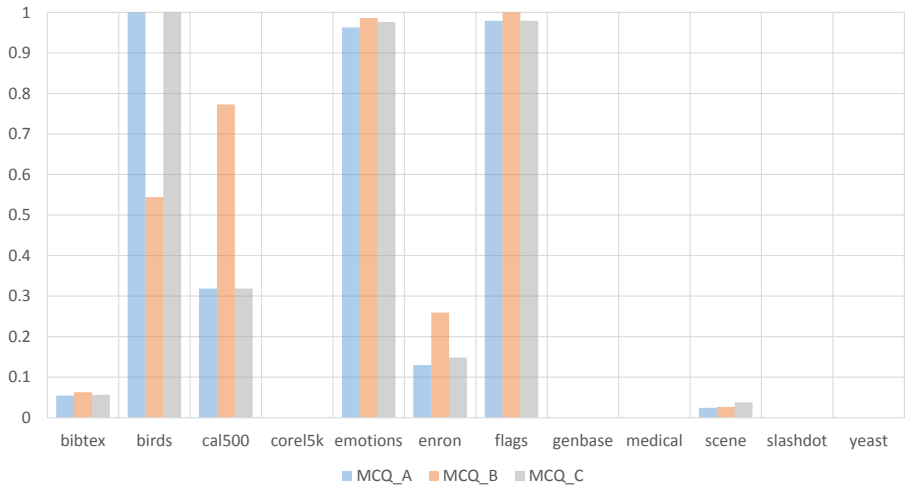

Figure 1: Consistency values obtained by the first group of data quality measures when using the HEOM distance to build the similarity relations.

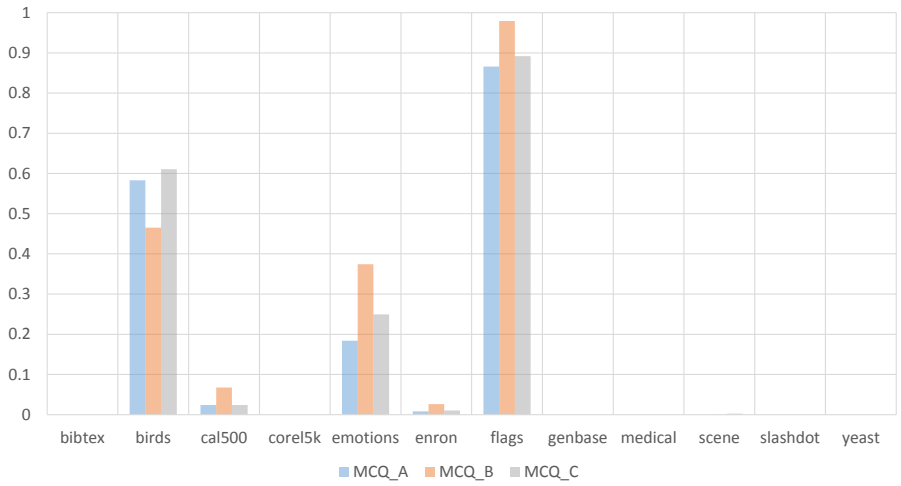

Figure 2: Consistency values obtained by the first group of data quality measures when using the HMOM distance to build the similarity relations.

4.4.1. Analyzing the impact of the threshold parameters

In the following experiment, we explore the impact of the similarity threshold parameters ξ_1 and ξ_2 on the consistency values calculated by the measures. By doing so, we report the average consistency values across all datasets for $\xi_1 \in$

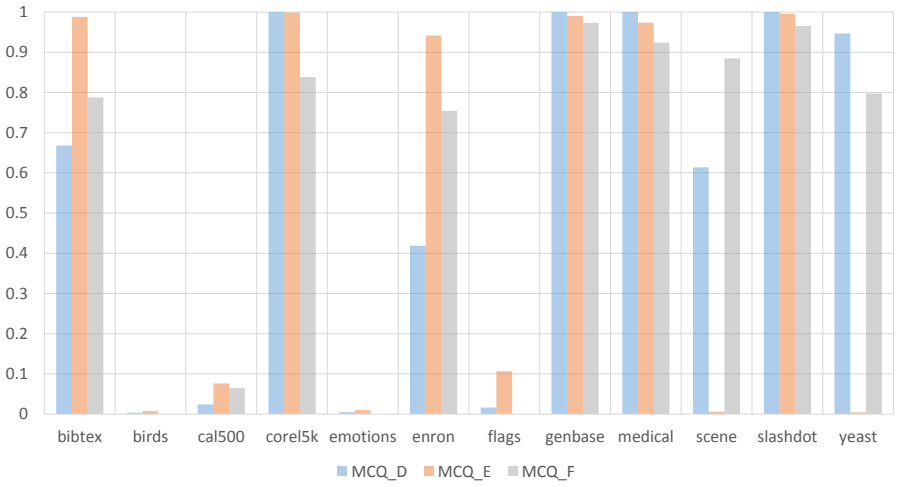

Figure 3: Consistency values obtained by the second group of data quality measures when using the HEOM distance to build the similarity relations.

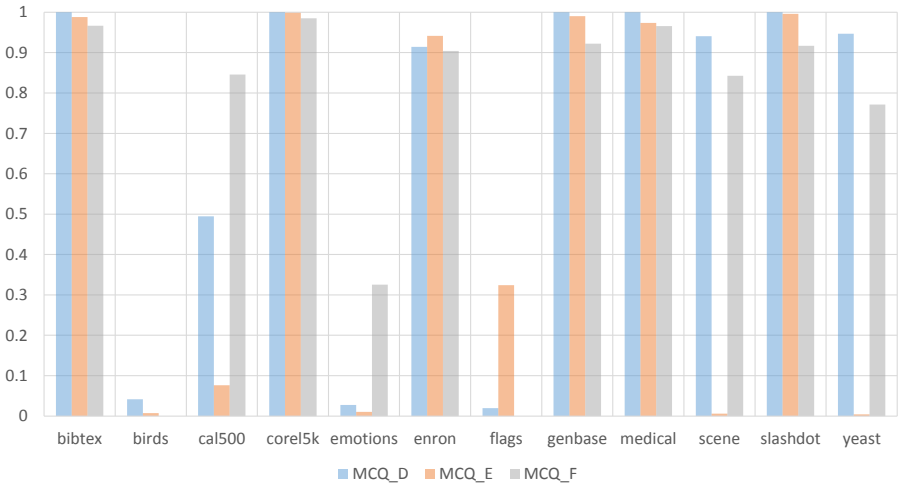

Figure 4: Consistency values obtained by the second group of data quality measures when using the HMOM distance to build the similarity relations.

$[0.5, 0.95]$ and $\xi_2 \in [0.3, 0.9]$ with step size 0.05.

Figure 5 shows the consistency values when changing the ξ_1 parameter for measures MCQ_A, MCQ_B and MCQ_C. The results indicate that the values obtained by measures MCQ_A, MCQ_B and MCQ_C are larger as ξ_1 approaches

1. This is somehow expected if we consider that larger threshold values lead to the subsets of indiscernible objects generated by the similarity relation containing fewer elements. Therefore, it is more likely to obtain a total inclusion of the object's similarity class in the decision class.

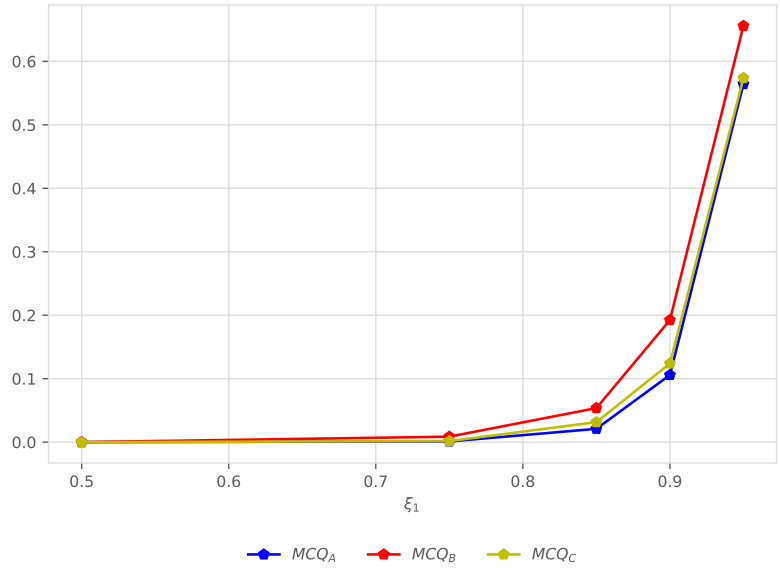

Figure 5: Average consistency values obtained with the first group of data quality measures for different values of the ξ_1 parameter in Equation (1). These measures report larger consistency values as the similarity threshold increases.

Figure 6 displays the consistency values when changing both ξ_1 and ξ_2 parameters for the MCQ_D measure. The results show that the MCQ_D measure reports larger values when $\xi_1 \in [0.5, 0.75]$ and $\xi_2 \in [0.3, 0.75]$. Unlike the first group of measures, MCQ_D attempts to reach a consensus between the granulation by condition and the granulation by decision. This means that we will obtain larger consistency values with more flexible thresholds. Note that, at the lower limits of the intervals, the largest values of the measures were reached, i.e. 0.5 and 0.3 for ξ_1 and ξ_2, respectively.

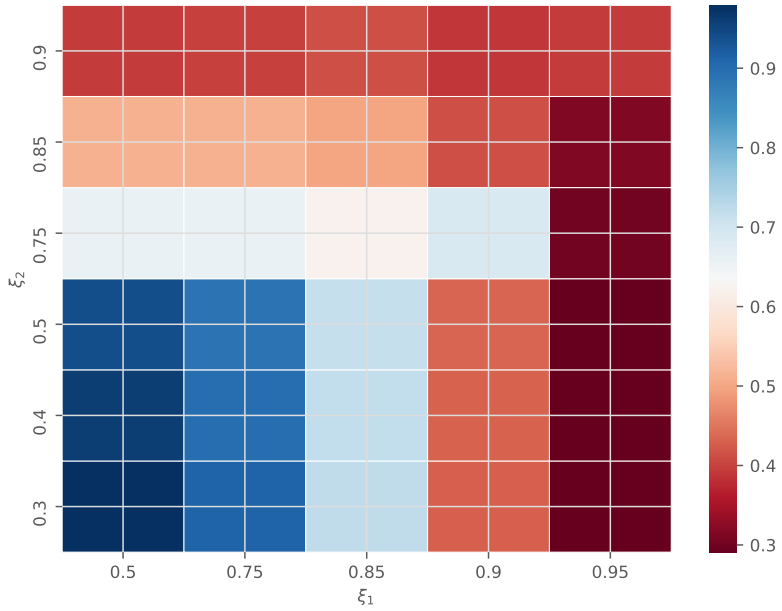

Figure 6: Average consistency values obtained with the MCQ_D measure for different values of the ξ_1 and ξ_2 parameters in Equations (1) and (9). This measure reports larger consistency values for smaller similarity threshold values.

It is worth mentioning that we did not carry out this simulation for MCQ_E and MCQ_F since these fuzzy measures can be effectively computed without specifying any similarity threshold parameter.

4.4.2. Analyzing the impact of the distance function

The analysis carried out at the beginning of this subsection indicated that, in some cases, the consistency values differ when changing the distance function, i.e. the HMOM distance often reports larger differences than HEOM. Therefore, we need to investigate whether the qualitative behavior of a certain measure holds when changing the distance function.

To conduct such a study, we transform numerical variables into symbolic representations with the aid of fuzzy variables. The states of these fuzzy variables are fuzzy sets associated with the following linguistic terms: *Very Low*, *Low*, *Medium*, *High*, and *Very High*. These linguistic terms provide a suitable

representation of the consistency values although more fuzzy sets will lead to more informative representations. Equations (23), (24), (25), (26), and (27) show the triangular membership functions for these fuzzy sets,

$$F_{VeryLow}(x) = \begin{cases} \frac{0.2-x}{0.2} & 0 \leq x \leq 0.2 \\ 0 & x \geq 0.2, \end{cases} \tag{23}$$

$$F_{Low}(x) = \begin{cases} 0 & x < 0.1 \\ \frac{x-0.1}{0.15} & 0.1 \leq x \leq 0.25 \\ \frac{0.4-x}{0.15} & 0.25 < x \leq 0.4 \\ 0 & x > 0.4, \end{cases} \tag{24}$$

$$F_{Medium}(x) = \begin{cases} 0 & x < 0.3 \\ \frac{x-0.3}{0.2} & 0.3 \leq x \leq 0.5 \\ \frac{0.7-x}{0.2} & 0.5 < x \leq 0.7 \\ 0 & x > 0.7, \end{cases} \tag{25}$$

$$F_{High}(x) = \begin{cases} 0 & x < 0.6 \\ \frac{x-0.6}{0.15} & 0.6 \leq x \leq 0.75 \\ \frac{0.9-x}{0.15} & 0.75 < x \leq 0.9 \\ 0 & x > 0.9, \end{cases} \tag{26}$$

$$F_{VeryHigh}(x) = \begin{cases} 0 & x < 0.8 \\ \\ \frac{x-0.8}{0.2} & 0.8 \leq x \leq 1. \end{cases} \tag{27}$$

Tables 4 and 5 show the symbolic consistency values for each measure using both the HEOM and HMOM distance functions, respectively. To obtain these symbolic values, we first evaluate each consistency value in each membership function and then apply the *principle of maximum membership* [39] which gives the linguistic term with the largest membership value. Notice that the maximum membership principle is equivalent to using a crisp partition with boundaries in the crossing points of membership functions.

Table 4: Symbolic consistency values when using the HEOM distance function.

DataSet	MCQ_A	MCQ_B	MCQ_C	MCQ_D	MCQ_E	MCQ_F
bibtex	Very Low	Very Low	Very Low	High	High	Very High
birds	Very High	Medium	Very High	Very Low	Very Low	Very Low
cal500	Low	High	Low	Very Low	Very Low	Very Low
corel5k	Very Low	Very Low	Very Low	Very High	High	Very High
emotions	Very High	Very High	Very High	Very Low	Very Low	Very Low
enron	Very Low	Low	Low	Medium	High	Very High
flags	Very High	Very High	Very High	Very Low	Very Low	Very Low
genbase	Very Low	Very Low	Very Low	Very High	Very High	Very High
medical	Very Low	Very Low	Very Low	Very High	Very High	Very High
scene	Very Low	Very Low	Very Low	Medium	Very High	Very Low
slashdot	Very Low	Very Low	Very Low	Very High	Very High	Very High
yeast	Very Low	Very Low	Very Low	Very High	High	Very Low

In order to explore whether or not there are significant differences between the values obtained by each measure when using different distance functions, we resorted to the Wilcoxon signed-rank test [40]. Table 6 reports the p-values computed with this test, the negative and positive ranks, and whether or not the null hypothesis was rejected.

Table 5: Symbolic consistency values when using the HMOM distance function.

DataSet	MCQ_A	MCQ_B	MCQ_C	MCQ_D	MCQ_E	MCQ_F
bibtex	Very Low	Very Low	Very Low	Very High	Very High	Very High
birds	Medium	Medium	Medium	Very Low	Very Low	Very Low
cal500	Very Low	Very Low	Very Low	Medium	High	Very Low
corel5k	Very Low	Very Low	Very Low	Very High	Very High	Very High
emotions	Low	Medium	Low	Very Low	Low	Very Low
enron	Very Low	Very Low	Very Low	Very High	Very High	Very High
flags	Very High	Very High	Very High	Very Low	Very Low	Low
genbase	Very Low	Very Low	Very Low	Very High	Very High	Very High
medical	Very Low	Very Low	Very Low	Very High	Very High	Very High
scene	Very Low	Very Low	Very Low	Very High	High	Very Low
slashdot	Very Low	Very Low	Very Low	Very High	Very High	Very High

Table 6: Results of the Wilcoxon signed rank test.

	p-value	Negative ranks	Positive ranks	Ties	Null hypothesis
$MCQ_A - HEOM$ vs $MCQ_A - HMOM$	0.1088	0	3	9	Not rejected
$MCQ_B - HEOM$ vs $MCQ_B - HMOM$	0.1088	0	3	9	Not rejected
$MCQ_C - HEOM$ vs $MCQ_C - HMOM$	0.0656	0	4	8	Not rejected
$MCQ_D - HEOM$ vs $MCQ_D - HMOM$	0.0587	4	0	8	Not rejected
$MCQ_E - HEOM$ vs $MCQ_E - HMOM$	0.0955	5	1	6	Not rejected
$MCQ_F - HEOM$ vs $MCQ_F - HMOM$	0.3173	1	0	11	Not rejected

For this experiment, the Wilcoxon test fails to reject the null hypothesis in each pairwise comparison (i.e., p-value > 0.05 for a 95% confidence interval). Therefore, we can conclude that the distance function does not significantly

affect the behavior of our measures.

4.5. Correlation analysis

This subsection analyses the correlation between the consistency values computed by the proposed data quality measures and a multi-label classifier's efficacy. In this experiment, we use the ML-kNN classifier for being a state-of-the-art lazy learner [9, 11]. In this case, we used the implementation available in MULAN [41] and the default parameter settings. Tables 7 and 8 show the consistency values reported by each measure and the HL values achieved by the ML-kNN algorithm across all datasets.

Table 7: HL and consistency values obtained using the HEOM distance.

DataSet	MCQ_A	MCQ_B	MCQ_C	MCQ_D	MCQ_E	MCQ_F	HL
bibtex	0.0549	0.0626	0.0565	0.6681	0.7880	0.9880	0.0136
birds	1.0000	0.5442	1.0000	0.0036	0.0000	0.0073	0.0472
cal500	0.3187	0.7729	0.3187	0.0241	0.0646	0.0763	0.1388
corel5k	0.0000	0.0000	0.0000	1.0000	0.8382	0.9987	0.0094
emotions	0.9629	0.9865	0.9764	0.0049	0.0000	0.0102	0.1951
enron	0.1298	0.2591	0.1486	0.4186	0.7544	0.9416	0.0523
flags	0.9794	1.0000	0.9794	0.0162	0.0000	0.1066	0.2536
genbase	0.0000	0.0000	0.0000	1.0000	0.9731	0.9903	0.0048
medical	0.0000	0.0000	0.0000	1.0000	0.9238	0.9737	0.0151
scene	0.0245	0.0266	0.0382	0.6136	0.8849	0.0059	0.0862
slashdot	0.0000	0.0000	0.0000	1.0000	0.9655	0.9960	0.0517
yeast	0.0000	0.0000	0.0000	0.9469	0.7975	0.0042	0.1933

To analyze the possible correlation between the estimated quality values and the performance values, we compute the Spearman correlation coefficient [42]. This measure quantifies the strength and direction of the monotonic association between two variables X and Y. Tables 9 and 10 show the correlation between the variable X (measure value) and Y (HL value) for each measure using HEOM and HMOM, respectively. In these tables, the first value represents the correlation coefficient while the second one reports the p-value for this test. The null

Table 8: HL and consistency values obtained using the HMOM distance.

DataSet	MCQ_A	MCQ_B	MCQ_C	MCQ_D	MCQ_E	MCQ_F	HL
bibtex	0.0000	0.0000	0.0000	1.0000	0.9664	0.9880	0.0136
birds	0.5829	0.4651	0.6109	0.0417	0.0000	0.0073	0.0472
cal500	0.0239	0.0677	0.0239	0.4947	0.8457	0.0765	0.1388
corel5k	0.0000	0.0000	0.0000	1.0000	0.9852	0.9987	0.0094
emotions	0.1838	0.3744	0.2496	0.0276	0.3255	0.0102	0.1951
enron	0.0082	0.0264	0.0106	0.9143	0.9042	0.9416	0.0523
flags	0.8660	0.9794	0.8918	0.0197	0.0000	0.3241	0.2536
genbase	0.0000	0.0000	0.0000	1.0000	0.9220	0.9903	0.0048
medical	0.0000	0.0000	0.0000	1.0000	0.9658	0.9737	0.0151
scene	0.0004	0.0004	0.0025	0.9406	0.8426	0.0060	0.0862
slashdot	0.0000	0.0000	0.0000	1.0000	0.9170	0.9960	0.0517
yeast	0.0000	0.0000	0.0000	0.9469	0.7713	0.0044	0.1933

hypothesis will be rejected (and it will be concluded that there is a monotonic correlation) when the p-value is lower than the level of significance (i.e., less than 0.05 if we adopt a 95% confidence interval). A negative correlation coefficient means that, for any two variables X and Y, an increase in X is associated with a decrease in Y. The intuition dictates that the HL values should decrease as the consistency values increase. In a nutshell, a large consistency value should serve as a strong indicator that the problem is easy to solve by a multi-label classifier, regardless of the TCS value.

The results indicate that there is a strong negative monotonic correlation between the values reported by MCQ_D, MCQ_E, and MCQ_F and the performance measure. Such a result aligns well with our hypothesis. This means that the consistency values are heavily related to the classifier's efficacy, such that larger consistency values often yield smaller HL values (i.e., larger prediction rates). More consistent datasets should be less difficult to solve since the data have less overlap among the classes, and thus, the machine learning algorithms are more likely to be effective. However, the measures MCQ_A,

Table 9: Correlation between HL and our measures by using the HEOM distance.

	Spearman Correlation	Sig.	Hypothesis
X: MCQ_A, Y: HL value	0.507580	0.092071	Not rejected
X: MCQ_B, Y: HL value	0.630849	0.027839	Rejected
X: MCQ_C, Y: HL value	0.507580	0.092071	Not rejected
X: MCQ_D, Y: HL value	-0.626776	0.029178	Rejected
X: MCQ_E, Y: HL value	-0.605649	0.036880	Rejected
X: MCQ_F, Y: HL value	-0.678322	0.015317	Rejected

Table 10: Correlation between HL and our measures by using the HMOM distance.

	Spearman Correlation	Sig.	Hypothesis
X: MCQ_A, Y: HL value	0.641961	0.024411	Rejected
X: MCQ_B, Y: HL value	0.641961	0.024411	Rejected
X: MCQ_C, Y: HL value	0.641961	0.024411	Rejected
X: MCQ_D, Y: HL value	-0.790374	0.002215	Rejected
X: MCQ_E, Y: HL value	-0.781087	0.002705	Rejected
X: MCQ_F, Y: HL value	-0.678322	0.015317	Rejected

MCQ_B, and MCQ_C do not align with our hypothesis since the negative correlation between the consistency values and the HL values is not evident. Therefore, the MCQ_D, MCQ_E, and MCQ_F measures are more suitable than the others to calculate the consistency value.

4.6. Comparison between MCQ_X and TCS

This subsection elaborates on the differences between our measures (mainly MCQ_D, MCQ_E, and MCQ_F) with the TCS measure since both are intended to estimate some characteristics of the dataset. While the TCS measure is an indicator of theoretical complexity based on structural characteristics (i.e., the number of input features, the number of labels, and the number of different label combinations), the proposed measures are intended to measure the consistency

of the information contained in the dataset.

While the TCS measure quantifies the problem size, our measures quantify the quality of the data describing the problem. Notice that a bigger dataset is not necessarily any more consistent or inconsistent. According to [1], the higher the TCS value, the more complex it should be to process the dataset, which can be reflected in a higher error in the learning process. Therefore, one would expect a directly proportional relationship between the TCS and HL values. The Spearman correlation test (see Table 11) indicates a negative monotonic correlation between TCS and HL measures.

Table 11: Correlation between HL and TCS measures.

	Spearman Correlation	Sig.	Hypothesis
X: TCS, Y: HL value	-0.699301	0.011374	Rejected

Although the TCS measure could be used as an estimator of the learning process's efficacy, it does not always provide consistent results. For instance, the datasets *medical* and *cal500* have a similar TCS values but report different errors during the learning process. More explicitly, the HL value obtained in the *medical* dataset is less than the one obtained in the *cal500* dataset, and however, its complexities are almost the same.

- *cal500* reports a TCS value of 15.59 (rather high), the HL value is 0.1388 (rather high), while the consistency values are $MCQ_D = 0.0241$, $MCQ_E = 0.0646$, and $MCQ_F = 0.0763$. The values of our measures and the HL value suggest that *cal500* is an inconsistent dataset.
- *medical* reports a TCS value of 15.62 (rather high), the HL value is 0.0151 (rather low), while the consistency values are MCQ_D=1, MCQ_E=0.9238, and MCQ_F=0.9737. The values of our measures and the HL value suggest that *medical* is a consistent dataset.

It should be mentioned that the above counterexample is not intended to neglect the relationship between the problem's structural complexity and the

learning error. Nevertheless, that relation is neither trivial nor straightforward. It goes without saying that, as often happens in the artificial intelligence field, there could be situations in which our measures could not capture the data quality as accurately as the TCS measure. But certainly, that is not what we have observed in the numerical simulations.

5. Concluding remarks

In this paper, we presented six data quality measures for MLC problems. Overall, the rationale behind the proposed measures consists of evaluating the quality of classification in situations in which an observation can be associated with multiple labels simultaneously. It should be noted that the proposed measures do not use any machine learning algorithm, so they are agnostic data quality measures. Therefore, the advantage of having such measures is that we can get insight into the expected complexity of solving a particular problem even before running any MLC algorithm.

The first three measures proposed in this paper (MCQ_A, MCQ_B, and MCQ_C) attempted to use the original *quality of classification* measure. To do that, we explored some definitions of what could be considered a decision class in a multi-label dataset. However, the attempt to directly adapt the *quality of the classification* measure to the multi-label scenario was rather unsuccessful. However, the remaining measures (MCQ_D, MCQ_E and MCQ_F) make it possible to estimate the complexity of the data at both global and local levels. It is worth mentioning that the local measures are particularly useful to identify which objects are difficult to classify. To derive these measures, we adapted the consistency concept to the multi-label setting. The numerical simulations for these measures show a strong negative correlation between the consistency values obtained by these three measures and the algorithm's performance. This result confirms that we do not need to build a classification model to estimate the problem's complexity beforehand.

The future research work will be oriented the connecting the proposed mea-

sures with the meta-learning field. The envisaged research includes obtaining "meta" rules providing guidelines for selecting multi-label classifiers when a new problem arises. Likewise, further strategies to lighten the computational complexity of our measures need to be explored.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable and constructive feedback.

[1] F. Herrera, F. Charte, A. J. Rivera, M. J. Del Jesus, Multilabel classification, in: Multilabel Classification, Springer, 2016, pp. 17–31.

[2] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Transactions on Knowledge and Data Engineering 26 (8) (2014) 1819–1837.

[3] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, Springer, 2009, pp. 667–685.

[4] F. Charte, A. J. Rivera, M. J. del Jesus, F. Herrera, Addressing imbalance in multilabel classification: Measures and random resampling algorithms, Neurocomputing 163 (2015) 3–16.

[5] F. Charte, A. Rivera, M. J. del Jesus, F. Herrera, On the impact of dataset complexity and sampling strategy in multilabel classifiers performance, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2016, pp. 500–511.

[6] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining (IJDWM) 3 (3) (2007) 1–13.

[7] S. Y. Sohn, Meta analysis of classification algorithms for pattern recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (11) (1999) 1137–1144.

[8] J.-R. Cano, Analysis of data complexity measures for classification, Expert Systems with Applications 40 (12) (2013) 4820–4831.

[9] T. K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Transactions on Pattern Analysis & Machine Intelligence (3) (2002) 289–300.

[10] P. Brazdil, C. G. Carrier, C. Soares, R. Vilalta, Metalearning: Applications to data mining, Springer Science & Business Media, 2008.

[11] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, T. K. Ho, How complex is your classification problem?: A survey on measuring classification complexity, ACM Computing Surveys (CSUR) 52 (5) (2019) 107.

[12] Z. Pawlak, Rough classification, International Journal of Man-Machine Studies 20 (5) (1984) 469–483.

[13] Z. Pawlak, Rough sets, International Journal of Computer & Information Sciences 11 (5) (1982) 341–356.

[14] Y. Filiberto, R. Bello, Y. Caballero, M. Frias, An analysis about the measure quality of similarity and its applications in machine learning, in: Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, Atlantis Press, 2013.

[15] Y. Caballero, R. Bello, L. Arco, M. García, E. Ramentol, Knowledge discovery using rough set theory, in: Advances in Machine Learning I, Springer, 2010, pp. 367–383.

[16] J. A. Sáez, J. Luengo, F. Herrera, Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification, Pattern Recognition 46 (1) (2013) 355–364.

[17] W. Pedrycz, A. Skowron, V. Kreinovich, Handbook of granular computing, John Wiley & Sons, 2008.

[18] Y. Yao, Information granulation and rough set approximation, International Journal of Intelligent Systems 16 (1) (2001) 87–104.

[19] R. Bello, J. L. Verdegay, Rough sets in the soft computing environment, Information Sciences 212 (2012) 1–14.

[20] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Transactions on Knowledge and Data Engineering 12 (2) (2000) 331–336.

[21] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (2-3) (1990) 191–209.

[22] W.-Z. Wu, J.-S. Mi, W.-X. Zhang, Generalized fuzzy rough sets, Information Sciences 151 (2003) 263–282.

[23] L. Coello, Y. Fernández, Y. Filiberto, R. Bello, Impact of weight initialization on multilayer perceptron using fuzzy similarity quality measure, in: Workshop on Engineering Applications, Springer, 2016, pp. 115–122.

[24] Y. Fernandez, L. Coello, Y. Filiberto, R. Bello, R. Falcon, Learning similarity measures from data with fuzzy sets and particle swarms, in: Electrical Engineering, Computing Science and Automatic Control (CCE), 2014 11th International Conference on, IEEE, 2014, pp. 1–6.

[25] D. Xu, Y. Tian, A comprehensive survey of clustering algorithms, Annals of Data Science 2 (2) (2015) 165–193.

[26] A. Nagpal, A. Jatain, D. Gaur, Review based on data clustering algorithms, in: 2013 IEEE Conference on Information & Communication Technologies, IEEE, 2013, pp. 298–303.

[27] L. C. Stoica, M. P. Cristescu, A.-M. R. Stancu, Making the k-means algorithm using the hamming distance, Group 1 (1) (2017) 1.

[28] S.-S. Choi, S.-H. Cha, C. C. Tappert, A survey of binary similarity and distance measures, Journal of Systemics, Cybernetics and Informatics 8 (1) (2010) 43–48.

[29] Y. Yao, N. Zhong, Potential applications of granular computing in knowledge discovery and data mining, in: Proceedings of World Multiconference on Systemics, Cybernetics and Informatics, Vol. 5, CiteSeerX, 1999, pp. 573–580.

[30] M. M. Deza, E. Deza, Encyclopedia of distances, in: Encyclopedia of Distances, Springer, 2009, pp. 1–583.

[31] L. P. Dinu, F. Manea, An efficient approach for the rank aggregation problem, Theoretical Computer Science 359 (1-3) (2006) 455–461.

[32] M. Diker, Textures and fuzzy unit operations in rough set theory: An approach to fuzzy rough set models, Fuzzy Sets and Systems 336 (2018) 27–53.

[33] M.-L. Zhang, Z.-H. Zhou, ML-kNN: A lazy learning approach to multi-label learning, Pattern Recognition 40 (7) (2007) 2038–2048.

[34] H. Liu, X. Wu, S. Zhang, Neighbor selection for multilabel classification, Neurocomputing 182 (2016) 187–196.

[35] G. Tsoumakas, E. Xioufis, J. Vilcek, I. Vlahavas, Mulan: multi-label dataset repository, URL http://mulan. sourceforge. net/datasets. html.

[36] F. Charte, D. Charte, A. Rivera, M. J. del Jesus, F. Herrera, R ultimate multilabel dataset repository, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2016, pp. 487–499.

[37] R. B. Pereira, A. Plastino, B. Zadrozny, L. H. Merschmann, Correlation analysis of performance measures for multi-label classification, Information Processing & Management 54 (3) (2018) 359–369.

[38] D. R. Wilson, T. R. Martinez, Improved heterogeneous distance functions, Journal of Artificial Intelligence Research 6 (1997) 1–34.

[39] C. Shi-quan, Fuzzy equivalence and multiobjective decision making, MM Gupta and T. Yamakawa, Elsevier Science Publishers (North-Holland).

[40] F. Wilcoxon, Individual comparisons by ranking methods, in: Breakthroughs in Statistics, Springer, 1992, pp. 196–202.

[41] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, Mulan: A java library for multi-label learning, Journal of Machine Learning Research 12 (Jul) (2011) 2411–2414.

[42] G. W. Corder, D. I. Foreman, Nonparametric statistics: A step-by-step approach, John Wiley & Sons, 2014.