Chapter 5
Effective Multi-Label Classification Using Data Preprocessing

Vaishali S. Tidake
https://orcid.org/0000-0003-4543-6361
MVPS's KBT College of Engineering, Nashik, India

Shirish S. Sane
K. K. Wagh Institute of Engineering Education and Research, Nashik, India

ABSTRACT

Usage of feature similarity is expected when the nearest neighbors are to be explored. Examples in multi-label datasets are associated with multiple labels. Hence, the use of label dissimilarity accompanied by feature similarity may reveal better neighbors. Information extracted from such neighbors is explored by devised MLFLD and MLFLD-MAXP algorithms. Among three distance metrics used for computation of label dissimilarity, Hamming distance has shown the most improved performance and hence used for further evaluation. The performance of implemented algorithms is compared with the state-of-the-art MLkNN algorithm. They showed an improvement for some datasets only. This chapter introduces parameters MLE and skew. MLE, skew, along with outlier parameter help to analyze multi-label and imbalanced nature of datasets. Investigation of datasets for various parameters and experimentation explored the need for data preprocessing for removing outliers. It revealed an improvement in the performance of implemented algorithms for all measures, and effectiveness is empirically validated.

INTRODUCTION

Many scenarios in the real-life today depict applications of multi-label data. A document may be related to health as well as education, according to its text. A piece of news may focus on new technology that is helpful for safety as well. An image may contain several objects like roads, shops, buildings, etc. Contents of a paper may be relevant to multiple domains. A video may focus on topics of networking

DOI: 10.4018/978-1-7998-7371-6.ch005

along with virtualization. Thus many objects reveal multiple semantic meanings. Many researchers are working for the last few decades on multi-label classification. It is a task that assigns with a thing a set of predefined labels as per its properties.

BACKGROUND

The related work about multi-label classification and label imbalance is presented here. For multi-label classification, there exist methods that use the *transformation* approach. It changes multi-label data such that methods for single-label classification can be used. Sometimes multi-label data is not modified. Thus *adaptation* methods modify the process of dealing with such data. There also exists an approach that ensembles multiple existing methods. CC (Read, 2009), MLkNN (Zhang & Zhou, 2007) and RAkEL (Tsoumakas et al., 2011) are examples of these three approaches respectively.

For few decades, many researchers have worked in the field of multi-label classification (Tsoumakas & Katakis, 2007) (Tsoumakas et al., 2009) (Trohidis et al., 2008) (Tsoumakas et al., 2010) (Madjarov et al., 2012) (Zhang & Zhou, 2014) (Tidake & Sane, 2018). K nearest neighbor has also been the choice of many researchers for multi-label classification. From the study, it is noticed that neighbors are obtained using only features always. In contrast, the scenario is different for data that is multi-label. Each instance belongs to a predefined set of labels. Hence it is possible to consider labels along with features for obtaining neighbors.

Zhang and Zhou discuss an approach in (Zhang & Zhou, 2007). It follows an *algorithm adaptation* approach. It is an improved version of the well-known nearest neighbor algorithm. Several researchers use it to perform multi-label classification. It utilizes feature similarity to determine nearest neighbors (Zhang & Zhou, 2005) (Zhang & Zhou, 2007) (Spyromitros-Xioufis et al., 2008). In the case of multi-label classification, since the instances are associated with multiple labels, label dissimilarity may also help determine a set of nearest neighbors.

Class imbalance also poses problems to multi-label classifiers and may lower their performance. According to Spyromitros-Xioufis (2011), label skew is considered a class imbalance when considering each class individually. Francisco et al. (2013) have proposed how to measure the level of imbalance in a multi-label scenario. They have also presented two dataset preprocessing methods specially designed for multi-label datasets. They used sampling and LP for preprocessing. Those label sets that occur in a majority (minority) were reduced (increased). A method was suggested by Huang et al. (2015) for the improvement of multi-label classifier involving several binary classifiers. It can be used for feature selection also. SOSHF was extended from structured forests (Zachary et al., 2017). At each node, it has used transformation followed by split action to tackle class imbalance. An imbalance ratio was defined using positive and negative samples (Zhang et al., 2018). This ratio and label correlation was considered to improve BR models. Liu and Tsoumakas (2018) have handled the imbalance faced by ECC. They used an ensemble of CC with random under-sampling that helps to balance the distribution of each class. COCOA method explored joint label correlation and imbalance ratio from skewness between positive and negative samples (Zhang et al., 2020). It induced an imbalanced multi-class classifier per label.

MAIN FOCUS OF THE CHAPTER

A novel algorithm adaptation approach called MLFLD (Sane & Tidake, 2020) considered features and labels of instances to determine nearest neighbors while assigning weights to the neighbors. When two instances possess similar features, the chances of its selection as the nearest neighbor is more. Though labels of these instances are different, the possibility of its choice as nearest neighbor is low. The experimentation presented has shown the importance of using both features and labels to improve the classifier's performance. It has also demonstrated how the usage of particular distance measure affected the performance of devised algorithms.

Datasets may have an imbalance in the form of feature values. That can be checked by examining the existence of outliers. At the same time, multi-label datasets may have an imbalance in the form of labels also. This imbalance was measured using MLE (multi-label examples), skew and outliers, among other characteristics. The first two parameters are introduced in this chapter. These parameters computed using experiments helped to analyze the multi-label and imbalanced nature of datasets. Datasets were preprocessed to remove outliers. The performance of algorithms before and after preprocessing was analyzed, keeping an eye on the dataset characteristics. There is a need to explore how to handle imbalance.

In the subsequent sections, the work adopted by authors for the handling of multi-label data is presented. Six variants of experiments for developed algorithms are also focused on. Then multi-label datasets and their properties are described. Next, experimental results are discussed, followed by a conclusion.

DEVISED PARAMETERS AND ALGORITHMS FOR MULTI-LABEL CLASSIFICATION

Before presenting the work adopted by authors to handle multi-label data, different general and introduced parameters for measuring the multi-label and imbalanced nature of datasets are shown in the current section. Then two devised algorithms are presented, followed by two conventional and one introduced distance measures used by algorithms.

Parameters to Measure Multi-Label and Imbalance Nature of Multi-Label Datasets

Along with general parameters, two introduced parameters helped to analyze the multi-label and imbalanced nature of datasets.

Let *AL* denotes the actual label set present in dataset *D*. Let *E* and *F* be numbers of examples and features in *D,* respectively, as in Table 1. A proposed parameter *MLE* denotes the number of Multi-Label Examples: those with a count of labels more than 1 (Eq. (1)). A more considerable value shows more multi-label examples.

$$MLE(D) = \frac{1}{|E|}\sum_{i=1}^{|E|} V(|AL_i| > 1). \quad (1)$$

Here *V (.) = 1* if a count of labels associated with instance *i* is more than 1, otherwise it is 0. Another proposed parameter is *skew* that denotes the proportion of the most frequent label set (Eq. (2)). A smaller value shows an imbalanced label set nature.

$$Skew(D) = \frac{1}{|E|} \max_{AL_i, AL_j \in D} \{AL_i | \mu(AL_i) > \mu(AL_j), \forall AL_j\}. \quad (2)$$

Here $\mu(x)$ .denotes occurrence count of label set *x* in dataset *D*. One more parameter used is an *outlier* that tells a number of features having std. deviation ±1.5 (3) from the mean (Eq. (3)). A larger value shows imbalanced nature in the form of feature values.

$$Outlier(D) = \frac{1}{|F|} \sum_{i=1}^{|F|} V(\tilde{A}_i \geq \pm 1.5). \quad (3)$$

Here *V (.) = 1* if the standard deviation of feature *i* is more than ± 1.5, else it is 0. Weka (Hall et al., 2009) and Mulan (Tsoumakas et al., 2011) libraries were used for computation of these parameters. Table 1 shows these parameters that give a glance at the multi-label and imbalanced nature of used datasets.

Algorithm MLFLD

An algorithm for Multi-Label classification by exploring Feature Similarity and Label Dissimilarity (MLFLD) was designed for selecting proper neighbors for improving the performance of a multi-label classifier. MLFLD took the following parameters as input: a multi-label dataset (*MLDB*) with *q* instances, threshold (*Th*), number of neighbors (*k*), smoothing factor (*p*), and the distance measure for label dissimilarity (*Ldistance*). It operated in two stages.

In stage one, prior probabilities of each label *c* were obtained using Eq. (4)-(5). $cnt^{(c)}$ .for label *c* was obtained from known instances.

$$P(H_c = 1) = \left(p + cnt^{(c)}\right) / (2 \times p + q). \quad (4)$$

$$P(H_c = 0) = 1 - P(H_c = 1). \quad (5)$$

Then MLFLD has used available labels of those instances that are already known. While searching for the neighbors, MLFLD utilized their features. Required data were obtained from these neighbors for each label and stored in $F_1^{(c)}[j]$.and $F_1^{(c)}[j]$.arrays. This information was utilized for the estimation of likelihood probabilities (Eq. (6)-(7)).

$$P\left(\mathrm{E}=\mathrm{j}|H_c=1\right)=\frac{\mathrm{p}+F_1^{(c)}[j]}{p\,x\left(1+\mathrm{k}\right)+\sum_{r=0}^{k}F_1^{(c)}[r]},0\le j\le k. \tag{6}$$

$$P\left(E=j|H_c=0\right)=\frac{p+F_0^{(c)}[j]}{p\,x\left(1+k\right)+\sum_{r=0}^{k}F_0^{(c)}[r]},0\le j\le k. \tag{7}$$

In stage two, estimated probabilities of label c were utilized to predict label c for an unlabeled instance using Eq. (8)-(9).

$$j=\sum_{m=1}^{k}N_m^{(c)}. \tag{8}$$

$$t_c=1,if\left(\frac{P\left(H_c=1\right)\times P\left(E=j|H_c=1\right)}{P\left(H_c=1\right)\times P\left(E=j|H_c=1\right)+P\left(H_c=0\right)\times P\left(E=j|H_c=0\right)}\right)\ge Th. \tag{9}$$

Algorithm MLFLD-MAXP

In most of the applications involving multi-label data, it is expected that an instance belongs to a minimum of one label (Read, 2010) (Godbole & Sarawagi, 2004) (Zhu et al., 2005) (Kiritchenko, 2005) (Ghamrawi & McCallum, 2005). Algorithm MLFLD was expanded to avoid the prediction of no label. Authors expanded algorithm MLFLD with MAXimum Probability (MLFLD-MAXP) that predicted the most probable label from the label set for an instance under consideration, using Eq. (10) (Tidake & Sane, 2021).

$$x=\arg max_c\left(\frac{P\left(H_c=1\right)\times P\left(E=j\mid H_c=1\right)}{P\left(H_c=1\right)\times P\left(E=j\mid H_c=1\right)+P\left(H_c=0\right)\times P\left(E=j\mid H_c=0\right)}\right). \tag{10}$$

Distance Metrics for Label Dissimilarity

From the study, it has been noticed that neighbors were obtained using features always. While the scenario for multi-label data is that each instance is relevant to a predefined set of labels. Hence both

devised algorithms have used labels and features of known instances to locate neighbors. They computed feature similarity using Euclidean distance (Han & Kamber, 2012) and label dissimilarity using distance measures, namely Hamming, Jaccard and SimIC as shown in Eq. (11)-(13).

Hamming distance obtains a difference between a total number of distinct and shared labels between the two instances (Read et al., 2008) (Godbole & Sarawagi, 2004). Jaccard distance (Han & Kamber, 2012) uses a ratio of intersection of labels to their union to compute distance (Pesquita et al., 2007) (Veloso et al., 2007). SimIC (Similarity of Information Content) is motivated from SimGIC distance (Aleksovski et al., 2009). It computed information for label *c* using its probability in the dataset.

$$\text{Hamming}\left(X_i, X_j\right) = \frac{\left|Labels\left(X_i\right) \cup Labels\left(X_j\right)\right| - \left|Labels\left(X_i\right) \cap Labels\left(X_j\right)\right|}{c}. \tag{11}$$

$$\text{Jaccard}\left(X_i, X_j\right) = 1 - \left(\frac{\left|Labels\left(X_i\right) \cap Labels\left(X_j\right)\right|}{\left|Labels\left(X_i\right) \cup Labels\left(X_j\right)\right|}\right). \tag{12}$$

$$\text{SimIC}\left(X_i, X_j\right) = 1 - \left(\frac{IC\left(Labels\left(X_i\right) \cap Labels\left(X_j\right)\right)}{IC\left(Labels\left(X_i\right) \cup Labels\left(X_j\right)\right)}\right). \tag{13}$$

For a set of labels $A = \left\{L_1, L_2 \ldots L_n\right\}$. $IC\left(A\right)$ .was obtained from the sum of the information content of $L_1, L_2 \ldots L_n$ .each from Eq. (14).

$$IC\left(c\right) = -\log\left(p\left(c\right)\right). \tag{14}$$

RESULTS AND DISCUSSION

Before discussion of results, an overview of used multi-label data is taken in the current section. Different values obtained through experiments for introduced parameters along with general characteristics of data are presented. The nature of datasets is analyzed based on these values. Then the performance of devised algorithms for six variants of distance measures is compared with MLkNN. Among six variants, the variant performing best is used in further data preprocessing experiments to analyze outliers' effect.

Multi-Label Data

Benchmark datasets are provided by resources such as Mulan (Tsoumakas et al., 2011) and MEKA (Read & Peter, 2012) (Tidake & Sane, 2016). Table 1 describes used multi-label datasets having numeric features only. All the datasets were normalized before use.

General characteristics of the benchmark datasets are shown in Table 1. Only the CAL500 dataset has labels approx. three times more than features. The rest of the datasets have feature count lesser or equal to label count. Also, it is essential to notice that in CAL500, each label set occurs precisely once. Hence the %Unique is 100.

Table 1. Characteristics of datasets

Datasets	Type	F	L	E	Cardinality	Density	% Unique	%Ex/ Label	% MLE	% Skew	% Outlier
Emotions	Media	72	6	593	1.868	0.311	4.6	31.0	70.0	13.7	18.9
Image	Media	294	5	2000	1.236	0.247	1.0	24.7	22.9	18.9	86.2
Scene	Media	294	6	2407	1.074	0.179	0.6	17.9	7.4	16.8	72.2
Yeast	Bio	103	14	2417	4.237	0.303	8.2	30.2	98.7	9.8	29.6
CAL500	Media	68	174	502	26.044	0.15	100.0	14.9	100	0.2	16.3
F: #Features, L: #Labels, E: #Examples											

Table 1 shows Label Cardinality and Label Density of datasets (Zhang & Zhou, 2007) (Carvalho & Freitas, 2009) (Read et al., 2009). They represent an average number of labels/example, and Cardinality/number of labels, respectively. Unique (some researchers denote it as label diversity) (Tsoumakas & Katakis, 2008) shows distinct combinations of labels present in the dataset.

From Figure 1(a), Emotions, Image and Scene, have Cardinality one. Many instances in them have only one label. In Yeast, Cardinality 4 shows many instances have approx. 4 labels. Only CAL500 has Cardinality 26, while the rest datasets have Cardinality less than five. All datasets have minimal Density, except Emotions and Yeast followed by Image. The first two datasets have around 30%, while the third dataset has about 25% labels associated with almost every example. Each label set in CAL500 occurs only once, which means its labelling scheme is very irregular than the remaining datasets.

From Figure 1(b), Scene and Image have only 7% and 22% records associated with more than one label, respectively. The remaining datasets contain more than 70% MLE.

%Skew shows that Scene and Image have higher label skew comparatively than that of Yeast and Emotions. More examples are associated with the most frequent label combination, whereas the remaining examples are associated with rare label combination. Skew in CAL500 is less.

Outliers deviate the performance of a classifier (M. Hall et al., 2009). From Table 1, both Image and Scene contain more %outliers shown by 86 and 72, respectively.

In Figure 2(a), %Skew shows conflicting performance than %Ex/Label. For more skew, %Ex/label is less and vice-versa. Scene and Image datasets have comparatively less unique and more skew label sets, as shown in Table 1. As in Figure 2(b), datasets contain 3-26 labels. But most datasets contain examples

Figure 1a. Label statistics

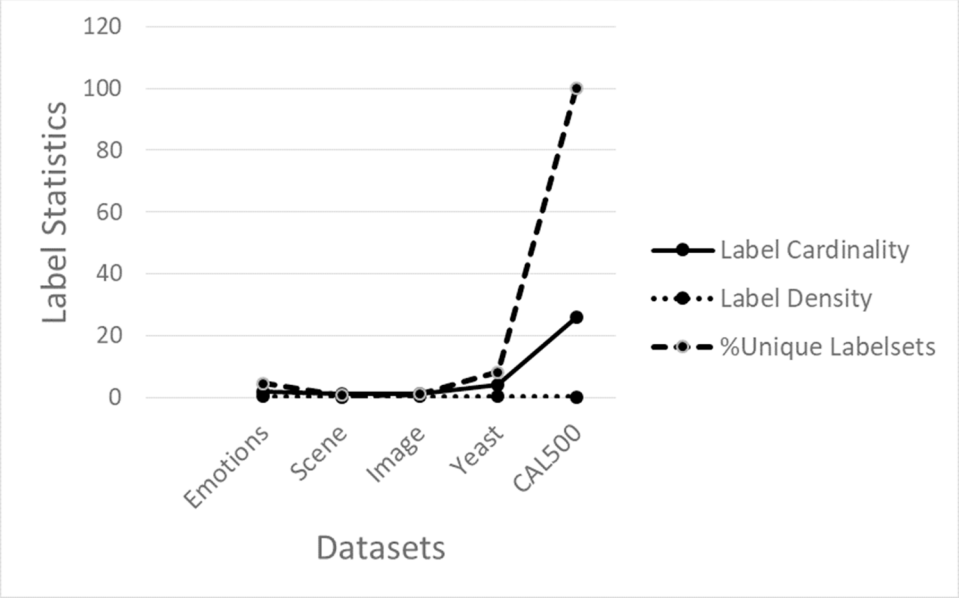


Figure 1b. Multi-label examples of datasets

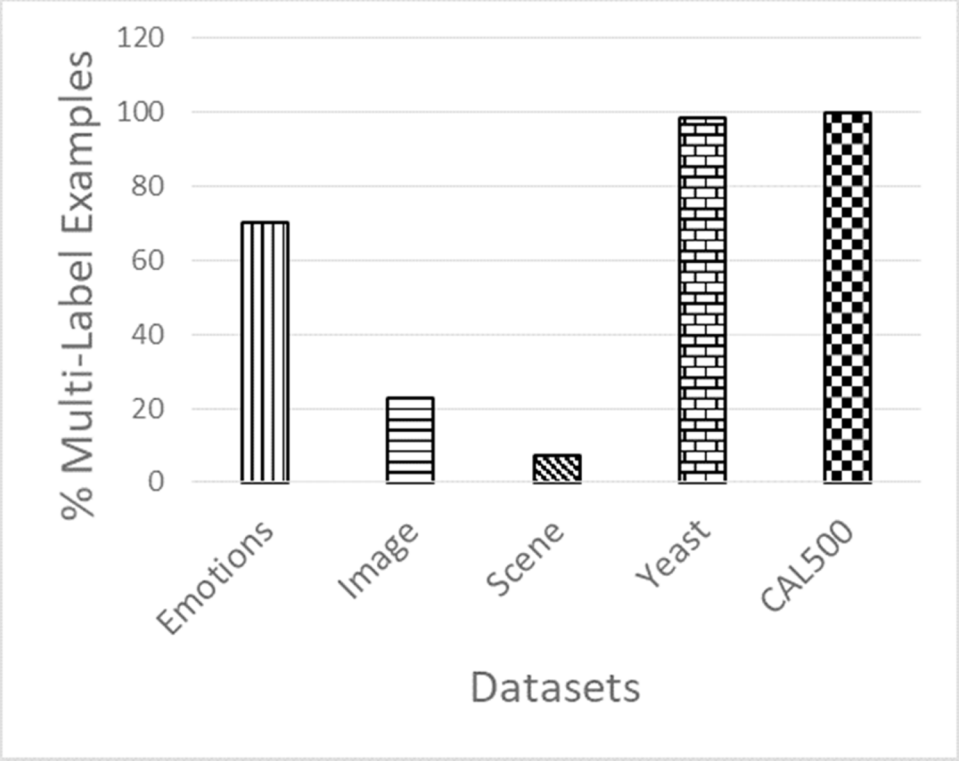
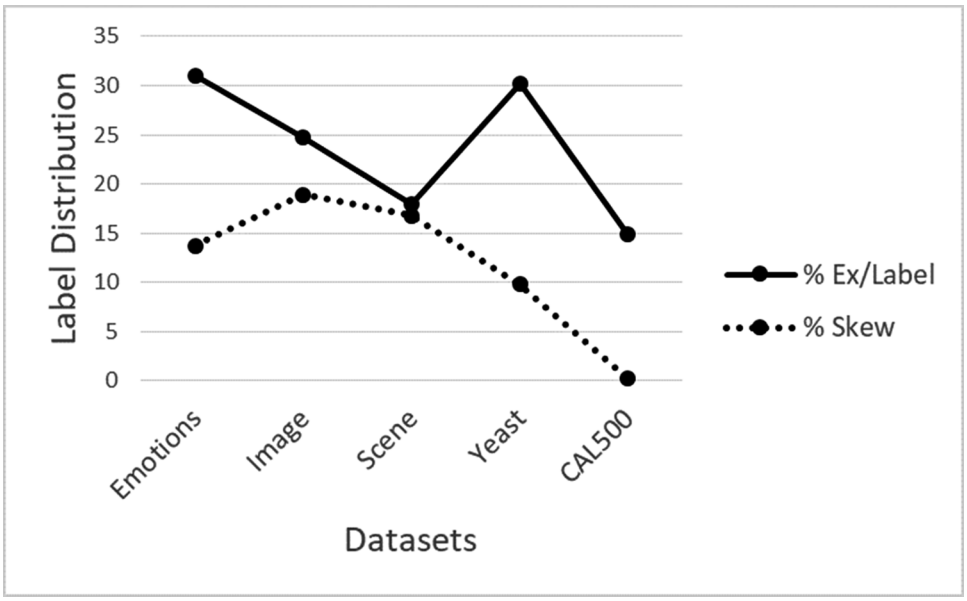


Figure 2a. Label distribution



associated with very few labels. Used datasets have 14-31% examples per label, but unique label sets are 1-8% only except CAL500.

Comparison of Performance for Label Dissimilarity Distance Measures

As three measures for label dissimilarity, namely Hamming, Jaccard and SimIC distance, were used, their effect is observed in this section. Ten criteria were used for performance evaluation as shown in Table 2 (Tsoumakas & Katakis, 2007) (Tsoumakas et al., 2009) (Trohidis et al., 2008) (Tsoumakas et al., 2010) (Madjarov et al., 2012) (Zhang & Zhou, 2014) (Tidake & Sane, 2018).

Figure 2b. Cardinality of labels

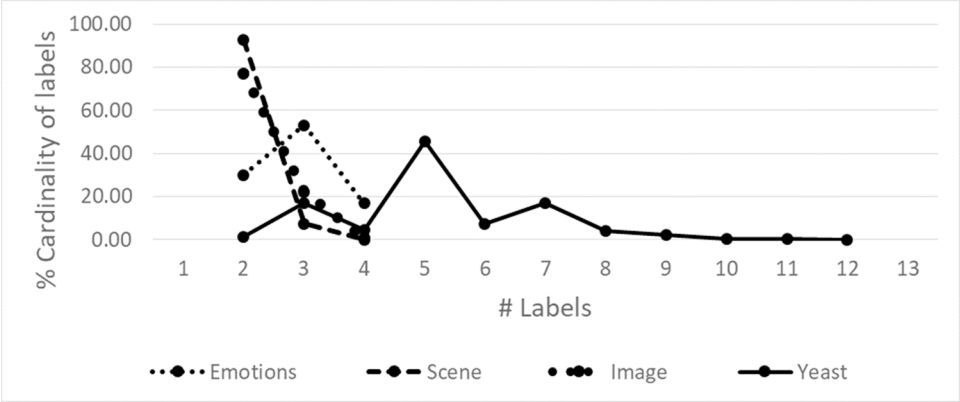


Table 2a. Effect of label dissimilarity on hamming loss (↓)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.1959	**0.1938**	0.1989	0.1952	**0.1938**	0.1952	0.1944
Image	0.1690	0.1631	0.1632	**0.1620**	0.1656	0.1661	0.1657
Scene	0.0861	0.0797	0.0795	**0.0792**	0.0812	0.0811	0.0807
Yeast	**0.1940**	0.1981	0.1967	0.2036	0.1977	0.1961	0.2041
CAL500	**0.1388**	0.1394	0.1393	0.1409	0.1394	0.1393	0.1409
Average	0.1568	**0.1548**	0.1555	0.1562	0.1555	0.1556	0.1572
Rank	6	1	2	5	3	4	7

Table 2b. Effect of label dissimilarity on ranking loss (↓)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.1959	**0.1938**	0.1989	0.1952	**0.1938**	0.1952	0.1944
Image	0.1690	0.1631	0.1632	**0.1620**	0.1656	0.1661	0.1657
Scene	0.0861	0.0797	0.0795	**0.0792**	0.0812	0.0811	0.0807
Yeast	**0.1940**	0.1981	0.1967	0.2036	0.1977	0.1961	0.2041
CAL500	**0.1388**	0.1394	0.1393	0.1409	0.1394	0.1393	0.1409
Average	0.1568	**0.1548**	0.1555	0.1562	0.1555	0.1556	0.1572
Rank	6	1	2	5	3	4	7

During the experimentations, the primary aim was to explore how measures of label dissimilarity influence MLFLD and MLFLD-MAXP. First, Euclidean and Hamming were used for feature similarity and label dissimilarity. Then Hamming was replaced by Jaccard and SimIC in further experiments. Obtained 6 variants were compared with each other and MLkNN.

Table 2c. Effect of label dissimilarity on one error (↓)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.2699	**0.2492**	0.2508	0.2610	**0.2492**	0.2508	0.2610
Image	0.3000	0.2916	0.2916	**0.2901**	0.2916	0.2916	0.2901
Scene	0.2256	0.2050	0.2050	**0.2046**	0.2050	0.2050	0.2046
Yeast	**0.2300**	0.2378	0.2311	0.2506	0.2378	0.2311	0.2506
CAL500	0.1176	0.1160	**0.1140**	0.1240	0.1160	**0.1140**	0.1240
Average	0.2286	0.2199	**0.2185**	0.2261	0.2199	**0.2185**	0.2261
Rank	7	3	1	5	3	1	5

Table 2d.Effect of label dissimilarity on coverage (↓)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	1.7764	**1.7102**	1.7542	1.7576	**1.7102**	1.7542	1.7576
Image	0.9390	**0.8964**	0.8964	0.8999	**0.8964**	0.8964	0.8999
Scene	0.4753	**0.4258**	0.4288	0.4304	**0.4258**	0.4288	0.4304
Yeast	**6.2750**	6.2905	6.3183	6.3697	6.2905	6.3183	6.3697
CAL500	130.564	130.524	**130.512**	130.652	130.524	**130.512**	130.652
Average	28.0059	**27.9694**	27.9819	28.0219	**27.9694**	27.9819	28.0219
Rank	5	1	3	6	1	3	6

Table 2e.Effect of label dissimilarity on average precision (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.8034	**0.8183**	0.8094	0.8061	**0.8183**	0.8094	0.8061
Image	0.8030	0.8105	**0.8106**	0.8104	0.8105	**0.8106**	0.8104
Scene	0.8652	0.8785	0.8785	0.8785	0.8785	0.8785	0.8785
Yeast	0.7650	0.7648	**0.7663**	0.7550	0.7648	**0.7663**	0.7550
CAL500	**0.4942**	0.4918	0.4927	0.4871	0.4915	0.4927	0.4871
Average	0.7462	**0.7528**	0.7515	0.7474	0.7527	0.7515	0.7474
Rank	7	1	3	5	2	3	5

In this section, the performance was studied for ten folds using six variants for label dissimilarity. It is detailed in Table 2(a)-2(k) with a summary at the end. For comparison, two criteria, namely minimum average rank and a maximum number of wins, were used.

Table 2f.Effect of label dissimilarity on accuracy (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.5340	0.5483	0.5158	0.5401	**0.5627**	0.5463	0.5619
Image	0.4937	0.5588	0.5709	0.5702	0.6169	**0.6187**	0.6179
Scene	0.6635	0.7083	0.7194	0.7110	0.7599	0.7604	**0.7615**
Yeast	0.5162	0.5116	0.5172	0.4862	0.5140	**0.5195**	0.4899
CAL500	0.1972	0.2023	0.1951	**0.2077**	0.2023	0.1951	**0.2077**
Average	0.4809	0.5059	0.5037	0.5030	**0.5312**	0.5280	0.5278
Rank	7	4	5	6	1	2	3

Table 2g. Effect of label dissimilarity on subset accuracy (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.2934	0.3051	0.2915	0.3068	0.3136	0.3017	**0.3169**
Image	0.4090	0.4632	0.4657	0.4702	**0.5108**	0.5063	0.5093
Scene	0.6248	0.6629	0.6758	0.6696	0.7117	0.7150	**0.7171**
Yeast	0.1874	0.2046	0.2033	0.1954	**0.2046**	0.2037	0.1959
CAL500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Average	0.3029	0.3272	0.3273	0.3284	**0.3481**	0.3453	0.3478
Rank	7	6	5	4	1	3	2

Table 2h. Effect of label dissimilarity on Ex-F1 (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.6141	0.6274	0.5901	0.6155	**0.6441**	0.6279	0.6415
Image	0.5223	0.5916	0.6070	0.6044	0.6532	**0.6572**	0.6551
Scene	0.6764	0.7235	0.7340	0.7249	0.7761	0.7756	**0.7763**
Yeast	**0.6204**	0.6109	0.6165	0.5819	0.6145	0.6201	0.5875
CAL500	0.3240	0.3311	0.3212	0.3377	0.3311	0.3212	**0.3377**
Average	0.5514	0.5769	0.5738	0.5729	**0.6038**	0.6004	0.5996
Rank	7	4	5	6	1	2	3

From Figure 3 and Table 2(k), <MLFLD-MAXP, Hamming, Euclidean> triplet topped among seven experiments with average rank 1.5 and 7 wins. To brief,

- All six variants got a better average rank than MLkNN. It showed 6.7 average rank and 0 wins.

Table 2i. Effect of label dissimilarity on macro-F1 (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.6226	0.6584	0.6399	0.6596	0.6609	0.6534	**0.6667**
Image	0.5815	0.6287	0.6358	0.6358	0.6482	**0.6507**	0.6496
Scene	0.7364	0.7683	0.7718	0.7696	**0.7795**	0.7789	0.7793
Yeast	**0.3853**	NaN	NaN	NaN	NaN	NaN	NaN
CAL500	**0.1714**	NaN	NaN	NaN	NaN	NaN	NaN
Average	0.4994	0.6851	0.6825	0.6883	0.6962	0.6943	**0.6985**
Rank	7	5	6	4	2	3	1

Table 2j. Effect of label dissimilarity on micro-F1 (↑)

Dataset	MLkNN	MLFLD			MAXP		
		Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Emotions	0.6610	0.6727	0.6476	0.6665	**0.6766**	0.6633	0.6745
Image	0.5842	0.6259	0.6346	0.6328	0.6449	**0.6483**	0.6461
Scene	0.7332	0.7617	0.7641	0.7621	0.7706	0.7702	**0.7709**
Yeast	0.6471	0.6426	0.6477	0.6218	0.6439	**0.6492**	0.6227
CAL500	0.3209	0.3294	0.3182	**0.3377**	0.3294	0.3182	**0.3377**
Average	0.5893	0.6065	0.6024	0.6042	**0.6131**	0.6098	0.6104
Rank	7	4	6	5	1	3	2

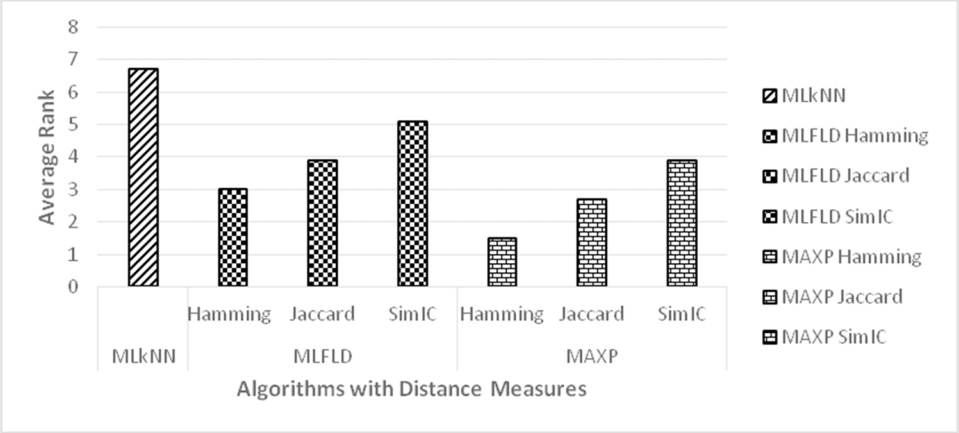
- For all metrics, variants of implemented algorithms exceeded MLkNN except hamming and ranking loss along with coverage.
- MLFLD and MLFLD-MAXP showed the same behavior for the first five measures.

From Table 1, Yeast and CAL500 were the most multi-label as implied by larger MLE values. But at the same time, their most frequent label set was associated with fewer examples indicated by smaller skew, inferring an imbalance label set. In contrast, these datasets have a lesser imbalance in terms of feature values. The emotions dataset also has a larger MLE. Simultaneously it has better skew and lesser outlier values. In Table 2(a)-2(k), MLkNN is better for datasets with very large MLE and smaller skew, while devised algorithms could not. They seemed evidenced better for datasets with lesser MLE but with more outliers.

Table 2k. Summarized performance for label dissimilarity

Performance Metric		MLkNN	MLFLD			MAXP		
			Hamming	Jaccard	SimIC	Hamming	Jaccard	SimIC
Hamming Loss	(↓)	0.1568	**0.1548**	0.1555	0.1562	0.1555	0.1556	0.1572
Ranking Loss	(↓)	0.1509	**0.1452**	0.1466	0.1494	**0.1452**	0.1466	0.1494
One Error	(↓)	0.2286	0.2199	**0.2185**	0.2261	0.2199	**0.2185**	0.2261
Coverage	(↓)	28.006	**27.969**	27.982	28.022	**27.969**	27.982	28.022
Avg. Precision	(↑)	0.7462	**0.7528**	0.7515	0.7474	**0.7528**	0.7515	0.7474
Accuracy	(↑)	0.4809	0.5059	0.5037	0.5030	**0.5312**	0.5280	0.5278
Subset Accuracy	(↑)	0.3029	0.3272	0.3273	0.3284	**0.3481**	0.3453	0.3478
Ex-F1	(↑)	0.5514	0.5769	0.5738	0.5729	**0.6038**	0.6004	0.5996
Macro-F1	(↑)	0.4994	0.6851	0.6825	0.6883	0.6962	0.6943	0.6985
Micro-F1	(↑)	0.5893	0.6065	0.6024	0.6042	**0.6131**	0.6098	0.6104
Exec. Time	(↓)	17	60	62	65	58	52	55
Avg. Rank	(↓)	6.7	3	3.9	5.1	**1.5**	2.7	3.9
#Wins	(↑)	0	4	1	0	**7**	1	1

Figure 3. Performance comparison for distance measures used for label dissimilarity



Performance After Outlier Removal

In the previous section, devised algorithms were seemed influenced mainly by the presence of more outliers. To check their influence, outliers were removed during preprocessing. The goal was to examine their performance on datasets of different nature. Here, devised algorithms were noticed for Hamming and Euclidean distance as their performance was seen to exceed compared to others. After removing outliers, datasets were supplied to three algorithms to be evaluated.

Summarized Table 3(b) has shown that both algorithms have defeated MLkNN after outlier removal from datasets. MLFLD-MAXP was seen much better than MLFLD. Figure 4(a)-4(e) has shown that for the first five metrics, devised algorithms behaved the same. For the rest five metrics, MLFLD-MAXP has surpassed MLFLD, as in Figure 4(g)-4(j). To summarize,

- Table 3 has illustrated that MLFLD always was better than MLFLD-MAXP for hamming loss enhancement, while MLFLD-MAXP appeared better after outlier removal.
- Both algorithms behaved the same for average precision, coverage, ranking loss, and one error (shown in Figure 4) with 2, 10, 33, and 37 percent increase than MLkNN, respectively. With MLFLD-MAXP and MLFLD, the highest enhancement was spotted as 46% and 35% for subset accuracy, and the same for accuracy was 32% and 24%, respectively. MLFLD-MAXP beat MLFLD for micro-F1 and ex-F1 with (21, 18) and (28, 21) percent, respectively. They have improved than MLkNN except for two datasets for macro-F1.
- The execution time was comparable for all experiments.

In Table 3(a)-3(b), after removing outliers, the scenario appeared very different. For all the metrics, both devised algorithms surpassed MLkNN. Thus, the imbalance of feature values influenced the designed algorithms for datasets with larger MLE and smaller label set skew.

Table 3a. Effect of outlier removal

(a) Hamming loss (↓)					*(b) Ranking loss* **(↓)**			
Dataset	**MLkNN**	**MLFLD**	**MAXP**		**Dataset**	**MLkNN**	**MLFLD**	**MAXP**
Emotions	0.1878	0.1115	**0.1104**		Emotions	0.1582	**0.0502**	**0.0502**
Scene	0.1052	0.0914	**0.0877**		Scene	0.0946	**0.0669**	**0.0669**
Image	0.1919	**0.1444**	0.1474		Image	0.2089	**0.1537**	**0.1537**
Yeast	0.1967	**0.1522**	**0.1522**		Yeast	0.1638	**0.0971**	**0.0971**
CAL500	0.1394	**0.1324**	**0.1324**		CAL500	0.1837	**0.1696**	**0.1696**
Average	0.1642	0.1264	**0.1260**		**Average**	0.1618	**0.1075**	**0.1075**
Rank	3	2	1		**Rank**	3	1	1
(c) One Error (↓)					*(d) Coverage* (↓)			
Dataset	**MLkNN**	**MLFLD**	**MAXP**		**Dataset**	**MLkNN**	**MLFLD**	**MAXP**
Emotions	0.2599	**0.1042**	**0.1042**		Emotions	1.7959	**1.1792**	**1.1792**
Scene	0.2910	**0.2302**	**0.2302**		Scene	0.5612	**0.4154**	**0.4154**
Image	0.3765	**0.2815**	**0.2815**		Image	1.0545	**0.8259**	**0.8259**
Yeast	0.2222	**0.1147**	**0.1147**		Yeast	6.2599	**5.1735**	**5.1735**
CAL500	0.1095	**0.0597**	**0.0597**		CAL500	131.057	**130.036**	**130.036**
Average	0.2518	**0.1581**	**0.1581**		**Average**	28.1457	**27.5260**	**27.5260**
Rank	3	1	1		**Rank**	3	1	1
(e) Average Precision (↑)					*(f) Accuracy* (↑)			
Dataset	**MLkNN**	**MLFLD**	**MAXP**		**Dataset**	**MLkNN**	**MLFLD**	**MAXP**
Emotions	0.8073	**0.9278**	**0.9278**		Emotions	0.5665	0.7276	**0.7380**
Scene	0.8301	**0.8700**	**0.8700**		Scene	0.6060	0.6667	**0.7407**
Image	0.7568	**0.8201**	**0.8201**		Image	0.3937	0.5722	**0.6630**
Yeast	0.7696	**0.8634**	**0.8634**		Yeast	0.5058	0.6235	**0.6236**
CAL500	0.4946	**0.5369**	**0.5369**		CAL500	0.1936	**0.2385**	**0.2385**
Average	0.7317	**0.8036**	**0.8036**		**Average**	0.4531	0.5657	**0.6008**
Rank	3	1	1		**Rank**	3	2	1
(g) Subset Accuracy (↑)					*(h) Ex-F1* (↑)			
Dataset	**MLkNN**	**MLFLD**	**MAXP**		**Dataset**	**MLkNN**	**MLFLD**	**MAXP**
Emotions	0.3223	0.5083	**0.5167**		Emotions	0.6458	0.7948	**0.8059**
Scene	0.5701	0.6189	**0.6907**		Scene	0.6179	0.6826	**0.7574**
Image	0.3501	0.5148	**0.5963**		Image	0.4084	0.5920	**0.6858**
Yeast	0.1805	**0.2806**	**0.2806**		Yeast	0.6111	0.7206	**0.7209**
CAL500	0.0000	0.0000	0.0000		CAL500	0.3186	**0.3781**	**0.3781**
Average	0.2846	0.3845	**0.4169**		**Average**	0.5204	0.6336	**0.6696**
Rank	3	2	1		**Rank**	3	2	1
(i) Macro-F1 (↑)					*(j) Micro-F1* (↑)			
Dataset	**MLkNN**	**MLFLD**	**MAXP**		**Dataset**	**MLkNN**	**MLFLD**	**MAXP**
Emotions	0.6404	0.8166	**0.8196**		Emotions	0.6814	0.8220	**0.8247**
Scene	0.6336	0.6998	**0.7397**		Scene	0.6715	0.7225	**0.7514**
Image	0.4455	0.5961	**0.6153**		Image	0.4768	0.6414	**0.6700**
Yeast	**0.3858**	NaN	NaN		Yeast	0.6396	0.7403	**0.7404**
CAL500	**0.1957**	NaN	NaN		CAL500	0.3147	**0.3831**	**0.3831**
Average	0.4602	0.7042	**0.7249**		**Average**	0.5568	0.6619	**0.6739**
Rank	3	2	1		**Rank**	3	2	1

FUTURE RESEARCH DIRECTIONS

In this work, multi-label data was observed for MLE, skew and outlier along with other properties. These were obtained through experimentation. It exhibited how performance was affected due to these properties. Multi-label data can be preprocessed further for a feature and instance selection or handling of skew nature. Observations of dataset characteristics showed that more MLE implied more skew. But

Table 3b. Summarized performance after outlier removal

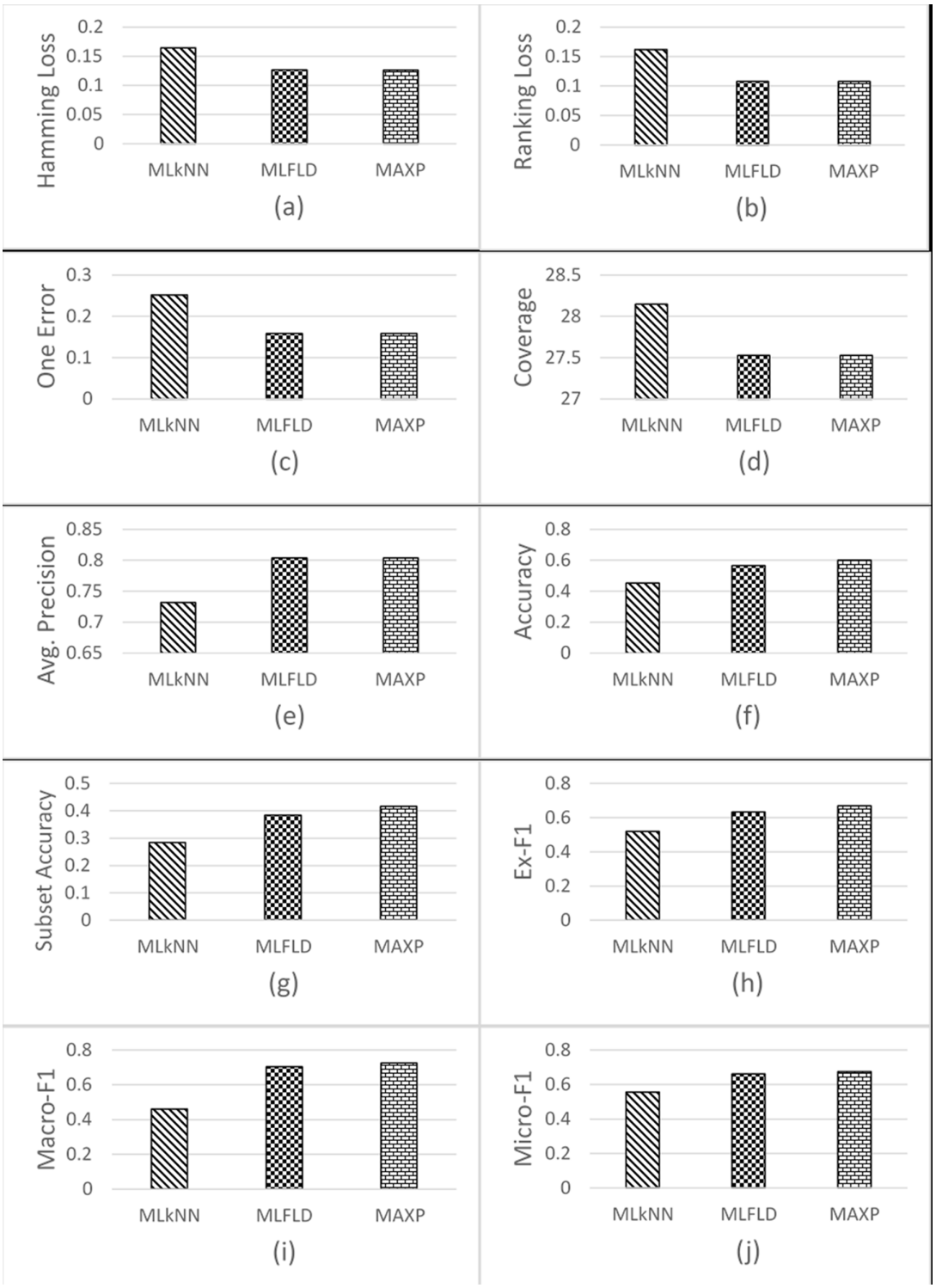
Performance Metric		MLkNN	MLFLD	MAXP
Hamming Loss	(↓)	0.1642	0.1264	**0.1260**
Ranking Loss	(↓)	0.1618	**0.1075**	**0.1075**
One Error	(↓)	0.2518	**0.1581**	**0.1581**
Coverage	(↓)	28.146	**27.526**	**27.526**
Avg. Precision	(↑)	0.7317	**0.8036**	**0.8036**
Accuracy	(↑)	0.4531	0.5657	**0.6008**
Subset Accuracy	(↑)	0.2846	0.3845	**0.4169**
Ex-F1	(↑)	0.5204	0.6336	**0.6696**
Macro-F1	(↑)	0.4602	0.7042	**0.7249**
Micro-F1	(↑)	0.5568	0.6619	**0.6739**
Exec. Time	(↓)	6	8	8
Avg. Rank	(↓)	3	1.6	**1**
#Wins	(↑)	0	4	**10**

at the same time, when MLE was less, outliers were more. It needs further investigation and empirical evaluation because experimentation was done on only five datasets.

CONCLUSION

Being associated with multiple labels, the use of label dissimilarity with feature similarity by MLFLD-MAXP has exceeded its performance. While the computation of label dissimilarity was observed for three distance measures, Hamming distance has shown maximum enhancement. When data was seen for outlier existence, its removal seemed more beneficial on MLFLD and MLFLD-MAXP. All the experiments implied that both the algorithms were sensitive to the presence of outliers. They were also affected by skew and the unique characteristics of datasets. It can be concluded that devised algorithms seemed more susceptible to datasets having very high MLE. The imbalance of feature values influenced the designed algorithms for datasets with larger MLE and smaller label set skew. Different forms of preprocessing on multi-label data can be further applied and observed.

Figure 4. Performance after outlier removal



REFERENCES

Aleksovski, D., Kocev, D., & Dzeroski, S. (2009). Evaluation of distance measures for hierarchical multilabel classification in functional genomics. *Proceedings of the 1st workshop on learning from multi-label data (MLD) held in conjunction with ECML/PKDD*, 5–16.

Charte, F., Rivera, A., del Jesus, M. J., & Herrera, F. (2013). A First Approach to Deal with Imbalance in Multi-label Datasets. HAIS 2013, LNAI 8073, 150–160.

Daniels, Z. A., & Metaxas, D. N. (2017). Addressing Imbalance in Multi-Label Classification Using Structured Hellinger Forests. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17).*

de Carvalho, A., & Freitas, A. A. (2009). A tutorial on multi-label classification techniques. In A. Abraham, A. E. Hassanien, & V. Snasel (Eds.), *Studies in Computational Intelligence 205* (pp. 177–195). Springer.

Ghamrawi, N., & McCallum, A. (2005) Collective multi-label classification. In *CIKM '05: 14th ACM International Conference on Information and Knowledge Management*. ACM Press. 10.1145/1099554.1099591

Godbole, S., & Sarawagi, S. (2004). *Discriminative methods for multi-labeled classification, Advances in Knowledge Discovery and Data Mining*. Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18. doi:10.1145/1656274.1656278

Han, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems.

Huang, J., Li, G.-R., Huang, Q.-M., & Wu, X.-D. (2015). Learning label specific features for multi-label classification. *Proc. IEEE Int. Conf. Data Min.*, 181–190. 10.1109/ICDM.2015.67

Kiritchenko, S. (2005). *Hierarchical Text Categorization and its Application to Bioinformatics* (PhD thesis). Queen's University, Kingston, Canada.

Liu, B., & Tsoumakas, G. (2018). *Making Classifier Chains Resilient to Class Imbalance*. ACML.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, *45*(9), 3084–3104. doi:10.1016/j.patcog.2012.03.004

Pesquita, C., Faria, D., Bastos, H., Falcao, A. O., & Couto, F. M. (2007). Evaluating GO-based Semantic Similarity Measures. *The 10th Annual Bio-Ontologies Meeting, ISMB/ECCB*.

Read, J. (2008). Multi-label classification using ensembles of pruned sets. *Proc. of 8th IEEE Int. Conf. on Data Mining*, 995-1000. 10.1109/ICDM.2008.74

Read, J. (2010). *Scalable Multi-label Classification*. The University of Waikato.

Read, J., & Peter, R. (2012). *MEKA: A multi-label extension to WEKA*. http://meka.sourceforge.net

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In *Proc. of European Conf. on Machine Learning and Knowledge Discovery in Databases: Part II. ECML PKDD '09*. Springer-Verlag. 10.1007/978-3-642-04174-7_17

Sane, S. S., & Tidake, V. S. (2020). Efficient Multi-label Classification using Attribute and Instance Selection. *Biosc. Biotech. Res. Comm. Special Issue*, *13*(14), 221–226. doi:10.21786/bbrc/13.14/52

Spyromitros-Xioufis, E. (2011). *Dealing with Concept Drift and Class Imbalance in Multi-label Stream Classification*. Thesis.

Spyromitros-Xioufis, E., Tsoumakas, G., & Vlahavas, I. (2008). An empirical study of lazy multi-label classification algorithms. *Proc. 5th Hellenic Conf. Artif. Intell.*, 401–406.

Tidake, V. S., & Sane, S. S. (2016). Multi-label learning with MEKA. *CSI Communications*, *2016*(August issue), 33–37.

Tidake, V. S., & Sane, S. S. (2018). Multi-label Classification: A Survey. *International Journal of Engineering and Technology, 7*(4.19), 1045-1054.

Tidake, V. S., & Sane, S. S. (2021). Effect of Distance Metrics on Multi-label Classification. In *Proceeding of First Doctoral Symposium on Natural Computing Research, Lecture Notes in Networks and Systems 169*. Springer Nature Singapore Pte Ltd.

Trohidis, K. (2008). Multi-Label Classification of Music into Emotions. *ISMIR, 8*.

Tsoumakas, G. (2010). *Mining multi-label data.* In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 667–686). Springer.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13. doi:10.4018/jdwm.2007070101

Tsoumakas, G., & Katakis, I. (2008). Effective and efficient multi-label classification in domains with large number of labels. *Proc. Work. Notes ECML PKDD Workshop MMD*.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-Labelsets for Multi-Label Classification. *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 1079–1089. doi:10.1109/TKDE.2010.164

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. (2011). MULAN: A Java library for multi-label learning. *Journal of Machine Learning Research*, *12*, 2411–2414.

Tsoumakas, G., Zhang, M. L., & Zhou, Z. H. (2009). *Tutorial on learning from multi-label data, in ECML PKDD*. Available: http://www.ecmlpkdd2009.net/wpcontent/uploads/2009/08/learning-from-multi-label-data.pdf

Veloso, A., Meira, W. Jr, Goncalves, M., & Zaki, M. (2007). Multi-label lazy associative classification. In *PKDD '07: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer

Zhang, M. L., Li, Y.-K., Liu, X.-Y., & Geng, X. (2018). Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, *12*(2), 191–202. doi:10.100711704-017-7031-7

Zhang, M. L., Li, Y.-K., Yang, H., & Liu, X.-Y. (2020). Towards Class-Imbalance Aware Multi-Label Learning. *IEEE Transactions on Cybernetics*, *2020*, 1–13. doi:10.1109/TCYB.2020.3027509 PMID:33206614

Zhang, M. L., & Zhou, Z. H. (2005). A k-nearest neighbor based algorithm for multi-label classification. *IEEE International Conference on Granular Computing*, 718-721. 10.1109/GRC.2005.1547385

Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. doi:10.1016/j.patcog.2006.12.019

Zhang, M. L., & Zhou, Z. H. (2014). A Review on Multi-Label Learning Algorithms. Knowledge and Data Engineering. *IEEE Transactions on.*, *26*(8), 1819–1837. doi:10.1109/TKDE.2013.39

Zhu, S., Ji, X., Xu, W., & Gong, Y. (2005). Multi-labelled classification using maximum entropy method. *SIGIR: '05: 27th Annual ACM Conference on Research and Development in Information Retrieval*, 274-281.