2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Confidence Factor and Feature Selection for Semi-supervised Multi-label Classification Methods

Fillipe M Rodrigues
Campus João Câmara
Federal Institute of Rio Grande do Norte (IFRN) BRAZIL
fillipe.rodrigues@ifrn.edu.br

Anne M P Canuto
Informatics and Applied Mathematics Department
Federal University of Rio Grande do Norte (UFRN) BRAZIL
anne@dimap.ufrn.br

Araken M Santos
Exact, Technology and Human Sciences Department
Federal Rural University of Semi-Arido (UFERSA) - BRAZIL
araken@ufersa.edu.br

Abstract— **In this paper, we investigate two important problems in multi-label classification algorithms, which are: the number of labeled instances and the high dimensionality of the labeled instances. In the literature, we can find several papers about multi-label classification problems, where an instance can be associated with more than one label simultaneously. One of the main issues with multi-label classification methods is that many of these require a high number of instances to be able to generalize in an efficient way. In order to solve this problem, we used semi-supervised learning, which combines labeled and unlabeled instances during the training process. In this sense, the semi-supervised learning may become an essential tool to define, efficiently, the process of automatic assignment of labels. Therefore, this paper presents four semi-supervised methods for the multi-label classification, focusing on the use of a confidence parameter in the process of automatic assignment of labels. In order to validate the feasibility of these methods, an empirical analysis will be conducted using high-dimensional datasets, aiming to evaluate the performance of such methods in different situations. In this case, we will apply a feature selection algorithm in order to reduce, in an efficient way, the number of features to be used by the classification methods.**

I. INTRODUCTION

In the single-label classification context, we have methods that assign each example (instance), to a single-label l from a finite set of disjoint labels L. In this case, a single label dataset D is composed of n examples $(x_1,l_1),(x_2,l_2),\ldots(x_n,l_n)$, where x represents the input data (set of attributes) and l is the representation of instances label [28]. Nevertheless, there are different application whose examples can be associated with a set of labels Y, where $Y \subseteq L$. In other words, each example could be associated with more than one label simultaneously. This type of application is called multi-label classification [1, 2]. Recently, researchers in different application domains has attracted attention to multi-label classification, such as in [3, 4, 5, 7, 9, 11, 12, 13], among others.

A natural limitation of the classification algorithms is that they need to have a set of labeled instances with a reasonable size in order to achieve a reasonable performance. However, the labeling process is often hard, expensive, and slow to obtain, because it may require human expertise. On the other hand, unlabelled instances are usually available in large quantity and they do not incur in high cost to collect. The problem with traditional classification algorithms is that they can not use unlabelled data in their training procedures [15].

This problem is particularly important in the multi-label context, since the number of possible combinations of the label attribute increases considerably. In this case, in multi-label tasks, the need of a large number of training examples is a critical problem. In view of the fact that the cost of manually labeling instances is very high and time-consuming process, researchers have been trying to smooth out this problem by using an automatic labeling process. One possible alternative for this is the use of semi-supervised learning. Basically, it uses information carried by the labeled instances in order to increase the performance of the classification models [16].

It is possible to find in the literature several applications of semi-supervised methods to single label classification such as in [17, 18, 19, 20]. However, few proposals have been developed in context of multi-label classification [21, 22, 15, 28]. Aiming is to add a contribution to this important subject, this paper applies three semi-supervised multi-label methods in high-dimensional problems, focusing on the use of a confidence parameter in the process of automatic assignment of labels.

This paper is an extension of the work done in [28], in which one confidence parameter was used in semi-supervised multi-label methods. In this paper, we apply a different approach in the process of automatic assignment of labels. Besides, this paper makes use of high-dimensional datasets, needing to apply a feature selection method. The overall idea of using confidence in the process of automatic assignment of labels is to minimize the inclusion of noisy, improving the overall classification accuracy. In order to analyze the performance of the proposed methods, an empirical analysis will be conducted, as well as a comparative analysis between the proposed methods and the methods proposed in [27, 28] (feature selection in high-dimensionality datasets). In this analysis, the proposed methods will be compared using different evaluation metrics. As a result of this, we aim to investigate the effect of the use of approach to calculate the confidence parameter

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

in the semi-supervised learning methods in the multi-label classification context.

II. MULTI-LABEL CLASSIFICATION

According to the number of labels that an example can be associated, a classification task can be divided into single or multi-label. Several methods have been proposed to be applied to multi-label classification problems, which can be broadly classified as algorithm adaptation and problem transformation methods [2, 4, 8]. In the first case, extensions of single-label classifiers have been proposed, adapting their internal mechanisms to allow them to be used in multi-label problems. Also, new algorithms can be developed specifically for multi-label problems [4]. Several algorithm adaptation methods are proposed in the literature, based on different algorithms, such as: lazy and associative methods, decision trees, support vector machines, probabilistic methods, neural networks, boosting, among others.

In problem transformation methods, a multi-label problem is transformed into a set of single-label classification problems. These methods are independent of the classification algorithms, since its operation does not depend on the classification algorithm. In the literature, different problem transformation methods have been proposed, such as in [25, 26]. In this paper, we will use four classification methods, which are:

- Binary Relevance (BR): In this method, the prediction of each label is considered as an independent binary classification task [2]. Therefore, BR builds *M* binary classifiers, one for each different label *L* (where *M* = *L*). For the classification of a new instance, it is considered the union of the labels l_i that are positively predicted by the *M* classifiers. The main disadvantage of BR is the fact that it assumes that the labels are assigned to an example in an independent way, ignoring all correlations that can exist among the assigned labels
- Label Powerset (LP): In this method, each possible combination of labels is defined as a label in a new single-label classification task [2]. For the recognition of a new instance, the single-label classifier of LP outputs the most likely label, which is actually a combination of labels. The main advantage of LP is that it takes into account the correlations among labels. However, the main drawback is the increasing complexity emerged from the large number of label subsets. Furthermore, the majority of these labels are associated with a small number of examples [2].
- Random k-labelsets (RA*k*EL): This method builds an ensemble of LP classifiers [6] and each LP classifier is trained using a small random subset of the combination of labels. An average decision is calculated for each label l_i in *L* and the final decision is positive for a given label if the average decision is larger than a given threshold *t*. It is important to highlight that RA*k*EL aims to take into account label correlations and, at the same time, to avoid the aforementioned problems of LP [6].
- Random k-labelsets *disjoint* (RA*k*ELd): it is an extension of Random k-labelsets (RA*k*EL) method [2]. This method uses the idea that the labelsets of each ensemble classifier has to be disjoint.

One of the main problems with multi-label applications is related to the evaluation of multi-label algorithms, since it requires the definition of new evaluation metrics. In a traditional single-layer problem, the most common evaluation metric is the accuracy level (or the error rate), in which it defines the number of patterns which were correctly classified. However, in a multi-label problem, a classifier can correctly assign an example to at least one of the labels it belongs to, but does not assign to all labels it belongs to. Also, a classifier could also assign an example to one or more labels it does not belong to [3].

In the literature it is possible to find some multi-label evaluation metrics. For the definition of these metrics, let a dataset of multi-label examples be denoted as (x_i, Y_i), $i=1, \ldots, N$, where $Y_i \subseteq L$ is the set of true labels and $L=\{\lambda_j : j=1 \ldots M\}$ is the set of all labels. Given an example x_i, the set of labels that are predicted by an multi-label method is denoted as Z_i, while $r_i(\lambda)$ is the ranking that predicted for a label λ. In this ranking, the most relevant label receives the highest rank (1), while the least relevant one receives the lowest rank (*M*) [2]. In this paper, four evaluation metrics will be used, which are:

- **Hamming Loss**: It defines the percentage of labels not predicted as well as incorrectly predicted labels. In other words, this metrics represents the number of examples that are associated with a wrong label or with a label not predicted. The smaller the value of hamming loss is, the better the performance is (the best performance when it is equal to 0). The equation for this metric is defined as follows.

$$\text{HammingLos} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \,\Delta\, Z_i|}{M} \quad \textbf{(1)}$$

- **Accuracy:** It symmetrically assesses how close Y_i *is to* Z_i.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad \textbf{(2)}$$

- **Subset Accuracy:** It represents the measure of accuracy limited way, considering as correct if all the labels predicted by one classifier are corrected.

$$\text{S Acc} = \frac{1}{N}\sum_{i=1}^{N} I\left(\left(|Z_i| - |Y_i|\right)\right) \quad \textbf{(3)}$$

- **F-Measure**: Its represents the combination of accuracy and recognition. This metric is the average harmonic of the two metrics Precision and Recall used as an aggregate performance score.

$$FM = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| - |Y_i|} \quad \textbf{(4)}$$

III. SEMI-SUPERVISED LEARNING

In the semi-supervised learning, it is possible to use partially-supervised information to guide to the learning process and increase the amount of evidence regarding the target problem. Therefore, the combination of semi-supervised learning with multi-label classification is the definition of efficient techniques to deal with problems where previously-classified data is very scarce. In general, semi-supervised learning uses both types of data, but mainly in situations where the number of labeled instances is small and the number of unlabeled instances is large [17, 18, 19].

Early studies in semi-supervised learning used methods with conceptual and algorithmic simplicity, such as expectation-maximization (EM)-based algorithms and self-training. About it, note [30]. As a consequence of self-training, the Co-training semi-supervised learning was proposed. Then, recently, graph-based semi-supervised learning methods have attracted great attention. Graph-based methods start with a graph where the nodes are the labeled and unlabelled instance, and (weighted) edges reflect the similarity of nodes. As the work reported here is the first attempt to incorporate semi-supervised learning in HML problems, we chose the self-training method as the semi-supervised learning method.

In this paper, we use a common and simple technique for semi-supervised learning that is known as self-training. In this method, a classifier uses its own predictions to teach itself [20]. The idea of self-training is firstly to create a classifier from a small number of label examples. After that, the underlying classifier is used to classify a proportion of unlabeled examples of the dataset. The underlying classifier is then retrained, including the newly labeled examples. This procedure is repeated until all the unlabeled instances have been moved to the labeled training set. For more details about this method, see [20].

The adaptation of the self-training method for multi-label problems is straightforward, allowing more than one label to be set to 1. In this case, a threshold is defined (usually 0.5) and label values higher than this threshold is set to 1 and values lower than this threshold is set to 0. In addition, its adaptation to hierarchical problems with the top-down approach is also straightforward, using in each node of the tree the semi-supervised approach to label the instances of the corresponding classes. Therefore, the adaptation of the self-training method for hierarchical multi-label problems uses the multi-label adaptation in each node of the tree used in the top-down approach of the hierarchical methods.

For a semi-supervised (SS) learning method, it is important to define two parameters, which are: the initial proportion of labeled instances and the proportion of instances to be labeled in one iteration. The idea is that a SS method uses the initial proportion of labeled instances to label, in a iteratively way, the instances of the unlabelled dataset.

IV. THE CONFIDENCE-BASED ALGORITHMS

One of the first studies about the combination of semi-supervised learning with multi-label methods was in [15] and four methods were proposed, which are:

- Semi-Supervised Binary Relevance (SSBR): It uses as kernel the Binary Relevance method (BR) to transform the multi-label data into single-label one. Then, we apply the labeling process using the self-training semi-supervised learning. At this stage, the set of labeled data is transformed into *M* subsets of data, one for each label of *L;*
- Semi-Supervised Label Powerset (SSLP): it uses the idea of Label Powerset (LP) to deal with multi-label data. In addition, as in the previous method, we use the same procedure of self-training to label all the unlabelled data;
- SSRAkEL: it applies the same steps to build an original Rakel ensemble of classifiers. The two algorithms (RAkEL and SSRAkEL) are different in the type of classifiers generated during the training process. While LP classifiers are generated in RAkEL, SSLP are generated in SSRAkEL.
- SSRAkELd: it is a disjoint labelset approach of SSRAkEL and uses the same steps to build an original RAkELd ensemble of classifiers. The two algorithms (RAkELd and SSRAkELd) are different in the type of classifiers generated during the training process

Nevertheless, we observed that the process of automatic assignment of labels was a difficult task. The main question is related to the random choice of the unlabeled examples to be labeled. In order to smooth out this problem, we propose an extension of the methods proposed in [15]. The main idea of these methods is to minimize the randomness in which the examples are chosen during the labeling process.

A. *Variability and Ranking-based Approach*

The first attempt to use confidence in semi-supervised multi-label methods was done in [28]. This technique can be described as follows.

1. Train a classifier C_i using the set of labeled data current;
2. Define and calculate a confidence factor of all unlabeled examples of the dataset;
3. Define a threshold and select only the instances whose confidence are higher than this threshold;
4. Rank the selected instances, in a descending order of confidence, the unlabeled examples;
5. Select the first *x* instances of this ranking;
6. Assign labels to all x instances:
   a. Set binary labels (0 or 1) for all possible labels of the problem, in which 1 is set to a label whose output is higher than 0.5 and 0 to outputs that are lower than 0.5;
7. Move the newly labeled examples for the set of labeled data.

The above process is repeated until the whole set of unlabeled data is empty. Therefore, in this approach, only examples whose output labels of the classifier with values

above a confidence threshold are taken into account. In this sense, a confidence factor, $0 \leq conf \leq 1$, will be used to control the process of automatically assigning labels in the semi-supervised learning process. In this paper, we will use a confidence threshold of 0.6.

In using this approach, we form four confidence-based methods using ranking, which are: SSBRcr (Semi-supervised BR with confidence using ranking), SSLPcr (Semi-supervised LP with confidence using ranking), SSRAkELcr (Semi-supervised RAkEL with confidence using ranking) and, the last one, the SSRAkELdcr (Semi-supervised RAkELd with confidence using ranking).

B. The proposed Approach

In [28], we observed that the use of a threshold in the confidence-based procedure caused a limitation in the number of instances to be labeled, since only few instances are selected at each iteration. This caused the automatic labeling processing slow and time consuming because it was necessary too many iterations in the semi-supervised learning. Because of this, we decided to propose a different approach in which no confidence threshold is used and this is proposed in this paper. This proposed approach is described in the following steps.

1. Train a classifier C_i using the set of labeled data current;
2. Calculate the confidence factor of all unlabeled examples of the dataset;
3. Rank, in a descending order of standard deviation, the unlabeled examples;
4. Select the first x examples of this ranking;
5. Assign a label (0 or 1) to all selected examples, according to the multi-label procedure;
6. Move the newly labeled examples for the set of labeled data.

The process is repeated until the whole set of unlabeled data is empty. Once again, we form four methods using ranking, which are: SSBRr (Semi-supervised BR using ranking), SSLPr (Semi-supervised LP using ranking), SSRAkELr (Semi-supervised RAkEL using ranking) and, the last one, the SSRAkELdr (Semi-supervised RAkELd using ranking).

The main difference between the first (variability and ranking) and second approach (variability) is the use of a confidence threshold in as the first criterion in the selection of unlabeled instances.

In step 2 of both approaches, a confidence factor is calculated. Any criterion could be used as long as it represents the importance of an unlabeled instance. In both approaches, we used the confidence (belongingness) degree that is provided by the multi-label classification methods for each class label. The main idea is to select examples in which the classification methods are certain that it belongs to some classes and not belong to others. Therefore, we need outputs with average confidence close to 0.5 and with high standard deviation. In this context, we then use the values obtained by calculation of standard deviations resultant of the confidences for each instance. In multi-label classification problems, each unlabelled instance presents three outputs: bipartition values, ranking values and confidences values. In this paper, we use the standard deviation equation using the confidence values produced by the unlabeled instance to all classes.

It is expected, therefore, that unlabeled examples with high value of the standard deviation have confidence degrees closer to the corresponding limits (close to 1 when the input patterns belongs to the class and close to 0 when it does not belong) than the unlabeled examples with low standard deviation value. For this reason, with rank all unlabeled examples in descending order of standard deviation (called confidence factor) and select the first x (proportion of unlabeled examples selected in each iteration). In the first approach, as we also use a confidence threshold, we can select less than x examples at each iteration.

V. Empirical Analysis

The main goal of this empirical analysis is to assess the performance of the proposed methods, comparing with their corresponding semi-supervised versions without confidence values. In this analysis, we tried to evaluate these methods in different application domains and using different evaluation metrics. In order to do this analysis, ten datasets of different domains will be used and all datasets are available at http://mlkd.csd.auth.gr/multilabel.html, which are described as follows:

- Emotions: It is related to the classification of songs according to the emotions they evoke[12];
- Genbase: It represents the biological data which are related to the functional classification of proteins [23].
- Medical: This dataset contains documents with a free-text summary of patient symptom histories and prognoses which are used to predict insurance codes [10, 24].
- Scene: This dataset is composed of semantic indexing of still scenes [5].
- Yeast: This dataset is related to protein functions classification [9] and it contains micro-array expressions and phylogenetic profiles for yeast genes.
- Bibtex: This Bibtex dataset uses a simple text file that can be created and modified using an arbitrary text editor format (dataset is available at http://cosmal.ucsd.edu/cal/projects/AnnRet/).
- CAL500: This dataset is composed of set of 1.700 musical notes from human that, generated, and describing 500 popular musical groups (dataset is available at www.bibtex.org).
- Corel5k: This dataset is represented by directory which contains used datasets to recognition object as Machine Translation, of Pinar Duygulu.
- Enron: This dataset represents a set of data that was collected and prepared by the CALO project – learning cognitive assistance and organizes.
- Mediamill: This dataset has resulted in studies on intelligent systems laboratory at the University of Amsterdam and aims to translate pixel to text for the purpose of image retrieval.

In Table I, these datasets are described in more details, describing the number of examples, the number of numeric (NUM) and discrete (DIS) attributes and the number of labels. It is also presented some multi-label data statistics, such as the number of distinct label subsets (DLS), the label cardinality (LC) and the label density (LD) [2]. Label cardinality represents the mean number of labels per example and label density is the same number divided by $|L|$.

TABLE I - STANDARD AND MULTILABEL STATISTICS

Datasets	Instance	Attributes		Lab	DLS	LC	LD
		NUM	DIS				
Emotions	593	72	0	6	27	1.868	0.311
Genbase	662	0	1.186	27	32	1.252	0.046
Medical	978	0	1.449	45	94	1.245	0.028
Scene	2712	294	0	6	15	1.074	0.179
Yeast	2417	103	0	14	198	4.327	0.302
Bibtex	7395	0	1.836	159	2.856	2.402	0.015
CAL500	502	68	9	174	502	26.044	0.150
Corel5k	5000	0	499	374	3.175	3.522	0.009
Enron	1702	0	1.001	53	753	3.378	0.064
Mediamill	43907	120	0	101	6.555	4.376	0.043

A. *Feature Section Method*

As can be observed in Table I, four datasets contains more than 1.000 attributes and one dataset contains almost 500 attributes. These datasets can be considered as large and a feature selection method is needed. In this paper, we apply a feature selection method, called RelieF, that has already been applied in multi-label problems in [27].

This method aims to search for features which provide good separability among classes and, at the same time, a reduction in the uncertainty inside the classes, respectively. Although the evaluation is done separately for each feature, ReliefF takes into consideration the effect of interacting features [14, 29].

The ReliefF algorithm calculated the quality of attributes of single-label data. The basic idea of ReliefF is to reward an attribute for having different values on a pair of similar instances of different classes, and punish it when different values on examples of the same class are found. For each feature, ReliefF outputs a value w, in the [-1, 1] interval, and the most important attributes have w values close to 1. For more information about RelieF, see [14, 27, 29].

B. *Methods and Methodology*

As already mentioned, we proposed the use of confidence in semi-supervised learning for multi-label classification problems, using two different approaches to select the unlabeled examples to be labeled.

As base classifiers to be used for all multi-label methods, we have chosen to use k-nearest neighbor (k-NN) classification method. We have done an initial analysis using several traditional classification algorithms and k-NN has delivered the best overall performance.

In semi-supervised methods, two main parameters play an important role, which are: the percentage of examples that were initially labeled and the proportion of unlabeled examples to be labeled in each iteration. For the first parameter, we chose 75% and, for the second parameter, we have chosen two possible values, 33.3%. In this case, six iterations will be needed to label all unlabeled examples. It is important to highlight that the setting of all parameters were done after an exhaustive empirical investigation.

The experimental results were evaluated using 4 evaluation measures, which are bipartition-based measures (Hamming Loss, F-Measure, Subset Accuracy and Accuracy).

All multi-label classification methods and supervised learning algorithms used in this work are implementations of the Weka-based [28] package for multi-label classification, called Mulan [24]. This package includes implementations of multi-label classification methods such as BR, LP, RAkEL and RAkELd. The implementations of the semi-supervised methods were obtained from adjustments made in Mulan, changing the training strategy for a semi-supervised strategy.

The experiments were conducted using the 10-fold cross-validation methodology. Therefore, all results presented in this paper refer to the mean over 10 different test sets. The parameters values used in the learning algorithms were suggested as default in Mulan.

VI. RESULTS AND DISCUSSION

The empirical analysis of this paper is divided into two parts, in which the first one presents the use of the proposed approach in all four semi-supervised multi-label methods, SSBRr, SSLPr, SSRAkELr and SSRAkELr. For this first part, the results will be analyzed using all ten datasets.

Table II shows the performance of SSBRr, SSLPr, SSRAkELr and SSRAkELdr. In this table, the results are presented using four different evaluation measures, described in Section II. The best result achieved by each semi-supervised method in each measure is represented by the shaded cells. It is important to highlight that, for some evaluation metrics, the best results were obtained with values close to 0, while, for other metrics, the best results were obtained with values close to 1. Therefore, along with the name of each metric, we added the symbols ↓ (the lowest means the best) and ↑ (the highest means the best) to represent the behavior of the evaluation metrics.

We can observe from Table II that the use of a confidence factor had a positive impact in the performance of the semi-supervised methods and SSLPr had obtained the best performance in the majority of cases, 19 cases out of 40. Among all proposed methods, the second approach (SSBRr) had a slightly better performance than the third one (SSRAkELdr), since it provided the highest performance in 11 cases (out of 40) while SSRAkELdr provided the best performance in 10 cases. The SSRAkELr achieve the best performance in only 4 cases (out of 40).

Among the proposed methods, we can observe that the SSRAkELr and SSRAkELdr had the best results in two datasets (Corel5k and Mediamill) and we applied the feature selection procedure in both datasets.

In analyzing Table II, we could observe SSBRr and SSLPr achieved positive results in datasets with no feature selection method, when compared with the results achieved in SSRAkELr and SSRAkELdr for the reduced datasets.

TABLE II - PERFORMANCE OF THE MULTI-LABEL METHODS WITH RELIABILITY PARAMETER

	SSBR*r*	SSLP*r*	SSRAkEL*r*	SSRAkELd*r*
Measure	**EMOTIONS**			
HLss↓	0.189	0.210	0.274	0.207
F-M↑	0.637	0.684	0.744	0.648
Acc↑	0.564	0.564	0.469	0.555
SAcc↑	0.274	0.309	0.293	0.320
	GENBASE			
HLss↓	0.005	0.005	0.021	0.014
F-M↑	0.955	0.973	0.515	0.950
Acc↑	0.915	0.946	0.515	0.950
SAcc↑	0.880	0.907	0.489	0.889
	MEDICAL			
HLss↓	0.022	0.036	0.041	0.038
F-M↑	0.490	0.721	0.677	0.671
Acc↑	0.655	0.593	0.378	0.498
SAcc↑	0.343	0.492	0.313	0.117
	SCENE			
HLss↓	0.091	0.096	0.107	0.111
F-M↑	0.667	0.717	0.721	0.733
Acc↑	0.662	0.812	0.744	0.702
SAcc↑	0.637	0.668	0.627	0.619
	YEAST			
HLss↓	0.205	0.232	0.222	0.209
F-M↑	0.572	0.634	0.439	0.590
Acc↑	0.555	0.466	0.390	0.478
SAcc↑	0.180	0.227	0.041	0.192
	BIBTEX			
HLss↓	0.015	0.021	0.030	0.014
F-M↑	0.031	0.218	0.093	0.066
Acc↑	0.029	0.179	0.044	0.066
SAcc↑	0.041	0.043	0.002	0.024
	CAL500			
HLss↓	0.145	0.197	0.148	0.199
F-M↑	0.349	0.339	0.264	0.337
Acc↑	0.217	0.211	0.155	0.205
SAcc↑	0.010	0.009	0.010	0.012
	COREL5K			
HLss↓	0.017	0.016	0.009	0.009
F-M↑	0.009	0.119	0.002	0.133
Acc↑	0.009	0.086	0.002	0.217
SAcc↑	0.022	0.024	0.019	0.015
	ENRON			
HLss↓	0.059	0.064	0.063	0.064
F-M↑	0.203	0.274	0.039	0.200
Acc↑	0.188	0.250	0.035	0.192
SAcc↑	0.088	0.120	0.081	0.083
	MEDIAMILL			
HLss↓	0.032	0.036	0.031	0.032
F-M↑	0.575	0.721	0.755	0.714
Acc↑	0.689	0.573	0.712	0.719
SAcc↑	0.155	0.210	0.049	0.158

A. *Comparative Analysis*

In the second part of this empirical analysis, the proposed methods are compared with some existing multi-label methods, proposed in [27, 28]. This comparative analysis is done using only two semi-supervised multi-label methods (SSBRr and SSLPr) and five datasets. This limitation is due to the fact that these methods and datasets are used in [27, 28] and, therefore, we can do a comparative analysis.

Tables III and IV present the results of all methods, for BR and LP, respectively. In these tables, for each dataset, the first lines represent the methods proposed in this paper, while the second lines represent the semi-supervised multi-label methods proposed in [28]. The main aim of this comparison is to investigate whether the approach to automatic label the unlabeled instances proposed in this paper provides better performance than the one proposed in [28]. In addition, the third lines represent the supervised methods proposed in [27] in which the same feature selection method is applied. The main aim of this comparison is to evaluate whether the feature selection method provides the best performance for the semi-supervised or supervised versions of the multi-label methods. Once again, the best result achieved by each method in each measure is represented by shaded cells.

As can be observed in Table III, the proposed SSBR (SSBRr) presents the best results in 8 cases, out of 20 analyzed cases. The best results for SSBRcr and RF-BR were in 6 and 7 cases, respectively. In other words, for the analyzed methods, we can state that the proposed method provided performance slightly better the other two methods. Therefore, it is possible to achieve satisfactory results in the proposed approach for BR.

TABLE III - PERFORMANCE OF THE MULTI-LABEL METHODS, FOR ALL VERSIONS OF BR

EMOTIONS				
	HLss↓	SAcc↑	Acc↑	F-M↑
SSBRr	0.189	0.273	0.564	0.622
SSBRcr	0.196	0.266	0.541	0.627
RF-BR	0.220	0.260	0.530	0.620
GENBASE				
SSBRr	0.005	0.938	0.981	0.989
SSBRcr	0.002	0.927	0.978	0.980
RF-BR	0.001	0.940	0.970	0.980
MEDICAL				
SSBRr	0.013	0.612	0.685	0.753
SSBRcr	0.017	0.633	0.407	0.408
RF-BR	0.010	0.640	0.710	0.730
SCENE				
SSBRr	0.090	0.669	0.662	0.667
SSBRcr	0.091	0.688	0.668	0.667
RF-BR	0.120	0.640	0.660	0.670
YEAST				
SSBRr	0.249	0.144	0.432	0.573
SSBRcr	0.195	0.147	0.524	0.608
RF-BR	0.240	0.120	0.480	0.590

Now, we analyze the performance of LP-based methods in Table IV. From this table, we can observe that the proposed method (SSLPr) had a slightly better performance than the other two (SSLPcr and RF-LP), since it provided the best results in 9 cases (out of 20), while SSLPcr provided the best performance in 7 cases out of 20. Finally the RF-LP achieved the best performance in only 4 cases (out of 20).

Based on Table IV, it is possible to analyze that the proposed approach had also had a positive impact in the performance of LP-based methods. In addition, the use of feature selection had a higher impact in the performance of the semi-supervised methods, since RF-LP had the poorest performance of all three analyzed methods.

B. Graphical Analysis

In order to support this comparative analysis, radar graphs based on the performance of all evaluation metrics used in this paper are plotted, where higher values indicate better performance for the following evaluation metrics (F-Measure, Accuracy, Subset Accuracy) and the lowest value indicating better performance for Hamming Loss. Thus, better results are the ones plotted far away from the center to F-Measure, Accuracy and Subset Accuracy and closer the center to Hamming Loss evaluation metric.

Figure 1 illustrates the radar graphic for BR-based methods. For simplicity reason, the radar graphs for only two datasets are illustrated, which are Scene and Yeast datasets. As can be observed, the radar graphics show, in general, a good performance for BR-based methods. In addition, the results were similar, but the proposed method (SSBRr) provided slightly better performance, especially for evaluation metrics Accuracy and Subset Accuracy.

FIGURE 1 – RADA GRAPHIC OF MULTI-LABEL METHODS BASED ON BR (SSBRR – SSBRCR – RF-BR)

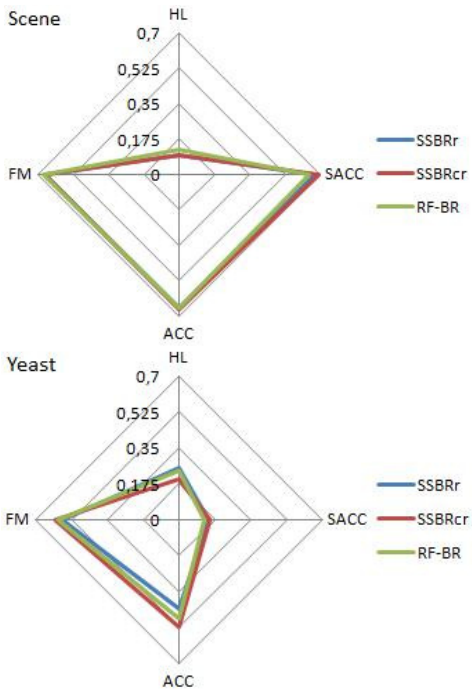
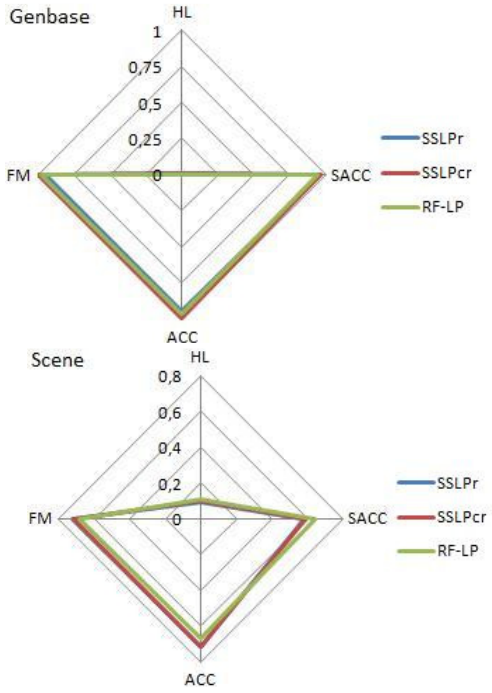


Figure 2 shows the radar graphs for the LP-based methods. Once again, we presented this graph for only two datasets, scene and Genbase. In general, once again, the radar graphs suggest a relative superiority of SSLPr, compared with the other two methods. The results of the radar graphs only confirms the results of Table III and IV, in which the proposed approach achieved slightly better performance than the other two analyzed methods.

TABLE IV - PERFORMANCE OF THE MULTI-LABEL METHODS, FOR ALL VERSIONS OF LP

	HLss↓	SAcc↑	Acc↑	F-M↑
EMOTIONS				
SSLPr	0.210	0.308	0.568	0.684
SSLPcr	0.214	0.299	0.551	0.668
RF-LP	0.220	0.280	0.540	0.630
GENBASE				
SSLPr	0.005	0.966	0.946	0.946
SSLPcr	0.005	0.967	0.996	0.995
RF-LP	0.001	0.940	0.970	0.980
MEDICAL				
SSLPr	0.019	0.544	0.594	0.610
SSLPcr	0.020	0.552	0.586	0.599
RF-LP	0.020	0.560	0.670	0.700
SCENE				
SSLPr	0.096	0.581	0.717	0.719
SSLPcr	0.097	0.583	0.712	0.714
RF-LP	0.110	0.640	0.670	0.680
YEAST				
SSLPr	0.220	0.148	0.517	0.623
SSLPcr	0.222	0.134	0.518	0.626
RF-LP	0.222	0.150	0.510	0.620

FIGURE 2 – GRAPHIC PERFORMANCE OF MULTI-LABEL METHODS BASED ON LP (SSLPR – SSLPCR – RF-LP)



VII. FINAL REMARKS

This paper proposed a confidence-based labeling process of unlabeled instances in semi-supervised learning approach in four multi-label classification methods that were originally

proposed in [15]. Since the proposed methods are extensions of existing semi-supervised methods, a comparative analysis was performed. These methods were applied to ten datasets, where we applied a feature selection method in high-dimensional datasets. In this empirical analysis, the analyzed methods were investigated using four different evaluation metrics. The results obtained in this analysis are very promising since the performance of the semi-supervised multi-label methods can be improved through the use of a confidence parameter in the labeling process.

The experimental analysis was divided into two parts in which the first one analyses the use of the proposed approach in all four semi-supervised multi-label methods. It was observed that the approach which achieved the best performance was SSLPr, with of 47.5% of the best cases, followed by SSBRr and SSRAkELdr, both methods with approximated percentage of 25% on the best cases.

In the second part of this empirical analysis, the proposed methods are compared with some existing multi-label methods, proposed in [27, 28]. As a result, we could observe that the confidence factor in semi-supervised learning proposed in this paper had positive effect, when compared with some existing multi-label methods. It was observed that the proposed approach achieved the best performance, providing the best results in 47.5% of the analyzed cases, for LP-based methods and in 40% of the analyzed cases for BR-based methods. Graphical statistical analysis was presented for a better understanding of the results between the approaches based on BR and LP in confidence-based labeling and feature selection methods [27, 28].

REFERENCES

[1] G. Tsoumakas. I. Katakis. and I. Vlahavas. “Effective and eficient multi-label classification in domains with large number of labels”. In: Proc. ECML/PKDD. Workshop on Mining Multidimensional Data. Antwerp. Belgium, 2008

[2] G. Tsoumakas. I. Katakis. and I. Vlahavas. “Mining Multi-label Data”. Data Mining and Knowledge Discovery Handbook. O. Maimon. L. Rokach (Ed.). Springer. 2nd edition. 2010.

[3] R. Cerri. R. R. Silva. and A. C. Carvalho.. “Comparing Methods for Multilabel Classification of Proteins Using Machine Learning Techniques”. BSB 2009. LNCS 5676. 109-120. 2009.

[4] J. Read. B. Pfahringer. G. Holmes. and E. Frank. “Classifier chains for multi-label classification”. Proc of Eur conf on Mach Learning and Knowledge Discovery in Databases. LNAI 5782(254-269). 2009.

[5] M. Boutell. J. Luo. X. Shen. and C. Brown. “Learning multi-label scene classification”. Pattern Recognition 37. 1757–1771. 2004.

[6] G. Tsoumakas. I. Katakis. and I. Vlahavas. “Random k-Labelsets for Multi-Label Classification”. IEEE T on Knowledge and Data Engineering. 2010.

[7] G. J. Qi. X. S. Hua. Y. Rui. J. Tang. T. Mei. and H. J. Zhang. “Correlative multi-label video annotation”. Proceedings of the 15th international conference on Multimedia. New York. NY. USA. 2007.

[8] G. Tsoumakas. I. Katakis. and I. Vlahavas. “Mining Multi-label Data”. Unpublished book chapter. 2009.

[9] A. Clare. and R. King. “Knowledge discovery in multi-label phenotype data”. Proc of the 5th Eur Conf on Principles of Data Mining and Knowledge Discovery. Freiburg. Germany . 2001.

[10] Computational medical center: Medical NLP challenge. http://www.computationalmedicine.org/challenge/index.php..

[11] Z. Barutcuoglu. R Schapire. and O Troyanskaya. “Hierarchical multi-label prediction of gene function”. Bioinformatics 22. 830–836. 2006.

[12] A. Wieczorkowska. P. Synak. and Z. Ras. “Multi-label classification of emotions in music”. Proc of the International Conference on Intelligent Information Processing and Web Mining . 307–315. 2006.

[13] K. Trohidis. G. Tsoumakas. G. Kalliris. and I. Vlahavas. “Multilabel classification of music into emotions”. Proc. 9th Int Conference on Music Information Retrieval. Philadelphia. PA. USA. 2008.

[14] N. Spolaôr. E. Cherman. and M. Monard. “Using ReliefF as feature selector in multi-label problems”(in portuguese). Centro LatinoAmericano de Estudos de Informática – CLEI, 2011.

[15] A.M. Santos and A.M.P. Canuto. "Using semi-supervised learning in multi-label classification problems," The 2012 International Joint Conference on Neural Networks (IJCNN), pp.1-8, 2012 doi: 10.1109/IJCNN.2012.6252800

[16] J. Metz, Jean and A. A. Freitas. “Extending hierarchical classification with semi-supervised learning”. UK Workshop on Computational Intelligence, 2009, Nottingham, UK. The 2009 UK Workshop on Computational Intelligence, 2009. p. 1-6.

[17] A. Singh, R. D. Nowak, and X. Zhu. “Unlabeled data: Now it helps, now it doesn't”. In NIPS, pages 1513-1520, 2008.

[18] A. Blum and T. Mitchell. “Combining labeled and unlabeled data with Co-Training”. In Proc. of The 11th Annual Conf. on Comp. Learning Theory, pages 92{100, New York, 1998. ACM Press.

[19] R. Jin, M. Wu, and R. Sukthankar.” Semisupervised collaborative text classification”. Proc. of the 18th European Conf. on Machine Learning, pages 600-607, Berlin, 2007. Springer-Verlag.

[20] X. Zhu. “Semi-supervised learning literature survey”. Technical Report 1530, Computer Sciences, University of Wisconsin- Madison, 2007. http://www.cs.wisc .edu/~jerryzhu/pub/ssl_survey.pdf.

[21] G Chen, Y Song, F Wang, C Zhang. “Semi-supervised Multi-label Learning by Solving a Sylvester Equation”. The 8th SIAM Conference on Data Mining (SDM), USA, pp. 410-419. 2008.

[22] Yi Liu, Rong Jin and Liu Yang. “Semi-supervised Multi-label Learning by Constrained Non-negative Matrix Factorization”. Proceedings of AAAI, 2006.

[23] D. Y. Hu and L. Reichel. “Krylov-subspace methods for the sylvester equation”. Linear Algebra and Its Applications, (172):283{313, 1992.

[24] G. Tsoumakas. R. Friberg. E. Spyromitros-Xiou. I. Katakis. and J. Vilcek. “Mulan software - java classes for multi-label classification Available at http://mlkd.csd.auth.gr/multilabel.html#Software or http://mulan.sourceforge.net

[25] D. Vilar. M. J. Castro and E. Sanchis. “Multi-Label Text Classification Using Multinomial Models.” Proc. Fourth Int'l Conf. España for Natural Language Processing (EsTAL '04). 2004.

[26] X. Luo. and N. A. Zincir-Heywood. “Evaluation of two systems on multi-class multi-label document classification” International Syposium on Methodologies for Intelligent Systems. 161–169. 2005.

[27] N. Spolaôr. E. Cherman. M. Monard. and H. Lee. “A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approache”. SciVerse ScienceDirec – Electronic Notes in Theoretical Computer Science 292 (2013) 135-151..

[28] F. Rodrigues. A. Santos. and A. Canuto. “Using Confidence Values in Multi-label Classification Problems with Semi-Supervised Learning”. IEEE-IJCNN, 2013.

[29] J. Demsar. “*Algorithms for subsetting attribute values with Relief* ” Machine Learning 78 (2010). pp. 421-428.

[30] O. Chapelle. “*Semi-Supervised Learning*” Machine Learning (2006). pp. 33-53.