Dimensionality Reduction in Multilabel Classification with Neural Networks

Jacek Mańdziuk
Faculty of Mathematics and Information Science
Warsaw University of Technology, Poland
mandziuk@mini.pw.edu.pl

Adam Żychowski
Faculty of Mathematics and Information Science
Warsaw University of Technology, Poland
a.zychowski@mini.pw.edu.pl

***Abstract*—A new neural network method for Dimensionality Reduction (DR) of the input feature space in Multilabel Classification (MC) problems is proposed and experimentally evaluated in this paper. The method (abbreviated as TCART-MR) can be used in two possible scenarios: either as a stand-alone DR pre-processing phase, preceding subsequent application of any particular MC algorithm, or as a compact MC approach in which TCART-MR is applied twice - first to DR task and then to MC problem with reduced input space.**

Extensive experimental results proved statistically relevant advantage of TCART-MR over three state-of-the-art approaches in DR domain (in the context of MC), as well as its superiority over 10 state-of-the-art MC algorithms listed in a recent MC survey paper. The MC tests were performed on a set of 9 benchmark problems and 16 evaluation measures (leading to 144 experimental cases in total).

I. Introduction

This paper considers a neural network model applicable to two research tasks: Multilabel Classification (MC) and Dimensionality Reduction (DR) of high-dimensional MC data. These two problems can be regarded either as one integrated research objective or treated in separation. Consequently, proposed Multilayer Perceptron (MLP) based method can be either applied to MC with or without initial DR phase (in the former case leading to superior results), or utilized for dimensionality reduction of MC data only, and subsequently followed by any other MC approach.

A. Multilabel classification

MC can be regarded as a natural extension of a standard (binary or multiclass) classification task. The main difference between *traditional* and MC classifiers lies in the size of the expected output of trained models. Instead of assigning one label (one class) to an object, in MC the goal is to assign a subset of all available labels. Formally, the problem of Multilabel Classification is defined as follows. Let us denote by $X = \{x_1, \dots, x_N\}, X \subseteq \mathbb{R}^n$ an instance space and by $Y = \{y_1, y_2, \dots, y_Q\}$ a finite set of Q predefined labels. The MC task consists in learning a function $h : X \to 2^Y$, which assigns a subset of corresponding labels to each object from X.

Accepted for the International Joint Conference on Neural Networks (IJCNN 2019) - Budapest, Hungary, July 2019

MC is commonly encountered in a wide range of real-life domains, including multimedia files categorization (e.g. music [1], video [2] or images [3]), text classification (e.g. automatic tagging of documents [4], articles [5] or e-mails [6]), or bioinformatics (e.g. finding probable diseases based on observed clinical symptoms [7] or discovering genomic functions [8]).

While the assignment of an ensemble of labels to each instance makes MC truly distinct from a standard, one label per instance, classification problem, at the same time both formulations have quite significant degree of commonality. One of the shared properties, which has a significant impact on the classification quality (accuracy) in both types of classification tasks is the *curse of dimensionality*. This property is related to the intrinsic inability of existing classification methods to deal with high-dimensional representation of the data, and is a well known impediment in building large, scalable classification system.

B. Dimensionality reduction

DR is a process of converting high-dimensional data into data of lower dimensionality while ensuring that the most pertinent information conveyed by this data is preserved. Generally speaking, there are two main approaches to DR: *feature selection* and *feature extraction*. Feature selection methods focus on finding a subset of original variables in order to remove irrelevant or redundant features. Feature extraction methods transform original data into smaller feature space usually with some information loss. Due to high practical relevance in domains of classification and regression, a wide range of DR methods were presented in the Machine Learning (ML) literature. The most popular ones are Principal Component Analysis (PCA) [9], Linear Discriminant Analysis (LDA) [10], Canonical Correlation Analysis (CCA) [11] or Non-negative Matrix Factorization (NMF) [12].

Majority of DR approaches, however, are not directly applicable to MC, due to their underlying assumption about having one output variable (one target class) per instance, while MC deals with subsets of output labels. Consequently, a bunch of other methods well suited to MC specificity was proposed. Some of them appropriately modify the baseline DR algorithms, for instance, Multilabel Linear Discriminant Analysis (MLDA) [13] is an MC extension of the LDA approach, or Conditional Principal Label Space Transformation (CPLST) [14] which adapts the CCA method for an MC setting. Other approaches are tailored specifically for an MC setup, e.g. the Multilabel Least Square (MLLS) [15] method which proposes a general framework for extraction of shared structures in MC. In [16], the Multilabel Dimensionality Reduction via Dependence Maximization (MDDM) algorithm is described which projects the original data onto a lower-dimensional feature space by maximizing the dependence between the original feature description and associated class labels, based on the Hilbert-Schmidt Independence Criterion. Multilabel Informed Latent Semantic Indexing (MLSI) proposed in [17] is an extension of a popular unsupervised Latent Semantic Indexing (LSI) [18] method by means of capturing correlations between multiple outputs (label sets). Also PCA can be directly applied to MC [19], as it does not assume any particular structure of the output values.

C. The main contribution

In this paper we address the problem of high dimensionality of MC data by extending our MLP based algorithm - *Tuned CARTesian-based Error Function for Multilabel Classification* (TCART-M) [20] - by means of its integration with an autonomous DR procedure that relies on weight examination of the trained TCART-M classifier. Extension of the TCART-M model by adding feature space dimensionality reduction phase leads to essentially new MC method and opens new research and application avenues of the model presented in [20]. This extended version of TCART-M will be call TCART-MR (TCART-M with DR phase) in the paper. It is worth underlining that, besides being an effective MC method itself, the TCART-MR algorithm can also be combined with any MC algorithm and serve as the initial step (reduction of input dimensionality) before the final MC takes place. In this respect, the applicability of TCART-MR extends beyond making an effective MC as the proposed method can also be applied directly for the purpose of DR (treated as a separate, stand-alone task) regardless of any particular MC method being subsequently applied. Both these aspects of TCART-MR application (MC and DR) are discussed and examined in detail in this paper.

In summary, the main contribution of this work is threefold:

- An improvement of the TCART-M approach to MC, initially proposed in [20], by adding an input dimensionality reduction phase. This extension of TCART-M is experimentally shown to significantly increase the quality of resulting classification and leads to an essentially novel, robust MC method (TCART-MR).
- An evaluation of a baseline TCART-M method and its newly-proposed extension (TCART-MR) in the task of MC based of their thorough comparison with 10 state-of-the-art approaches to MC, based on 9 benchmark problems and 16 error measures.
- An experimental verification of DR efficacy of TCART-MR by means of its direct comparison with 3 state-of-the-art DR methods widely utilized in MC domain.

The remainder of this paper is arranged as follows. Section II presents an overview of the state-of-the-art comparative approaches to MC, which are used for the experimental evaluation of the proposed method. Section III provides a description of the base TCART-M algorithm and introduces its extension TCART-MR that enhances TCART-M by autonomous input dimensionality reduction procedure. The next section presents the experimental setup: method parametrization, benchmark sets and evaluation measures, followed by a presentation of experimental results in Section V. The paper is concluded in the last section.

II. State-of-the-art approaches to the MC task

A wide range of practical applications stimulated and inspired research development in the field of MC, most notably in the last 15 years. In effect, a multitude of approaches to MC have been developed and published in the literature. Despite the variety of underlying ideas, all of them can roughly be divided into 3 categories: *problem transformation methods*, *algorithm adaptation methods* and *ensemble methods*.

In the rest of this section 9 methods (each of them belonging to one of these three classes) which are recognized as state-of-the-art algorithms by well-established multilabel survey paper [21], and are used as reference points for the TCART-M and TCART-MR assessment in the experimental section, are briefly characterized.

Problem transformation methods refer to algorithms which transform the problem of MC into other well-established learning scenario. For instance, *Binary relevance* (BR) [3] which is one of the most popular and, at the same time, one of the simplest algorithms relies on decomposing the multilabel learning problem into Q independent binary classification tasks. Each of these binary classification problems corresponds to one label in the label space (one vs. all classification). This transformed classification setting is subsequently approached with any binary classification method. Another example of a problem transformation method is the *Classifier chaining* algorithm (CC) [22]. Similarly to BR, it also uses binary classifiers, but not in the form of an ensemble but aligned in a sequence (chain of classifiers). Another popular method of this type is *Calibrated label ranking* (CLR) [23] which transforms the multilabel learning problem into a *label ranking* task, in which ranking the labels associated to a given sample is performed by means of a sequence of pairwise comparisons [24].

The second category of MC algorithms, *algorithm adaptation methods*, contains approaches which adapt one of the popular machine learning techniques (decision trees, k nearest neighbors, neural networks, etc.) to make it suitable for dealing with the MC task formulation. In particular, in *Multilabel k-nearest neighbors* (ML-kNN) [25], for each test instance, its k nearest neighbors among the training samples (with known label sets) are first identified. Next, based on these label sets, the *maximum a posteriori* (MAP) estimation is calculated and used to determine the set of labels for the considered sample. Another example is the *Backpropagation for Multilabel Learning* (BP-MLL) approach [26] which employs a one-hidden-layer MLP architecture with the input and output layers corresponding to the dimensionality of the data and the size of the set of possible labels (Q), respectively. BP-MLL is trained with the backpropagation algorithm with suitably defined error function. This method was an inspiration for our baseline TCART-M approach and is discussed in more detail in Section III.

The last group, *ensemble methods*, consists of algorithms which tackle the MC problem by independently running multiple instances of classifiers (e.g. some of those mentioned above) and combining their results by means of a voting scheme. Generally speaking, the approaches belonging to this category are the most powerful, but at the same time, the most time-consuming and difficult to parameterize MC methods [21]. A prominent example of this class is the *Ensembles of Classifier Chains* (ECC) [22] method which uses multiple instances of the CC algorithm, described above. Another ensemble method, *RAndom k-labELsets* (RAkEL) [27], randomly creates subsets of k labels and train separately a label powerset classifier for each of them. The final assignment of labels for a given test instance is defined independently for each label by a voting procedure involving all classifiers containing this label. Yet another algorithm from this group, *Hierarchy Of Multilabel classifiERs* (HOMER) [28], which is designated for large multilabel data sets, groups similar labels into subsets by means of a balanced clustering algorithm similar to k-Means, and then applies another instance of an MC algorithm to solve each of these smaller problems. The fourth method from this group considered in our experiments is *Random Forest of Predictive Clustering Trees* (RF-PCT) [21] which uses Predictive Clustering Trees [29] as a baseline classifier for randomly sampled training subsets.

Several deep learning neural network methods were also proposed, however, their applicability is usually limited to a certain domain or class of problems, e.g. text classification [30], [31], X-Ray images classification [32], health risk prediction [33], or images annotation [34].

III. TCART-M AND TCART-MR METHODS

TCART-M method is implemented as a one-hidden-layer (1hl) MLP with a suitably designed error function and the output layer of size Q (the number of possible labels). The base formulation of the method (denoted by CART-M) was inspired by the *Backpropagation for Multilabel Learning* (BP-MLL) model [26]. In particular, both BP-MLL and CART-M networks use the 1hl MLP architecture and are trained with backpropagation algorithm. The error function of BP-MLL is of the following form:

$$E_{BP-MLL} = \sum_{p=1}^{m} \frac{\sum_{(r,s)\in Y_p \times \overline{Y}_p} e^{-(c_r^p - c_s^p)}}{|Y_p||\overline{Y_p}|} \tag{1}$$

where m denotes the number of training instances, c_q^p is a network's output of the neuron associated with the q-th label in response to the p-th training sample (x_p), $Y_p \subseteq Y$ is a set of labels assigned to that p-th sample, and $\overline{Y_p} = Y \setminus Y_p$. For a given input sample x_p, minimization of E_{BP-MLL} leads to higher output values of neurons associated with labels that belong to Y_p (correct ones) than with those which does not belong to the p-th label set (incorrect ones). The chosen form of the error function (1) is closely related to the *ranking loss* criterion.

Once the training is completed, the set of labels $h(x) \subseteq Y$ assigned to a given test instance x is defined as follows [26]:

$$h(x) = \{q \in Y : c_q(x) > t(x)\} \tag{2}$$

where $c_q(x)$ denotes the q-th output in response to input x and $t(x)$ is a threshold associated with x. For each possible threshold value, its potential application result (the number of correctly assigned labels) is computed. Afterwards, the threshold with the best result is selected. Please note that a set of threshold values which lead to different results is finite (e.g. $\{0, c_1(x), \ldots, c_Q(x)\}$) and only these candidate thresholds need to be checked.

The above formulation of the error function was enhanced in [35] in the three following ways. Firstly, *integration of the threshold value into the error function* (and the training process) was proposed. Secondly, an *independent threshold for each label* was considered, so as to make the system more flexible and consequently more effective. Finally, comparisons within *any pair of output values* representing classes belonging to $Y_p \times \overline{Y}_p$, and their respective threshold values were taken into account. In effect, the error function introduced in [35] for the *CARTesian-based Error Function for Multilabel Classification* (CART-M) model was of the following form:

$$\begin{aligned} &E_{CART-M} = \\ &\sum_{p=1}^{m} \Bigg(\frac{\sum\limits_{(r,s)\in Y_p \times \overline{Y}_p} \left(e^{-(c_{2r}^p - c_{2s}^p)} + e^{-(c_{2s+1}^p - c_{2r+1}^p)}\right)}{2|Y_p||\overline{Y}_p| + |Y_p|^2 + |\overline{Y}_p|^2} \\ &+ \frac{\sum\limits_{r\in Y_p} \sum\limits_{t\in Y_p} e^{-(c_{2r}^p - c_{2t+1}^p)} + \sum\limits_{s\in \overline{Y}_p} \sum\limits_{t\in \overline{Y}_p} e^{-(c_{2t+1}^p - c_{2s}^p)}}{2|Y_p||\overline{Y}_p| + |Y_p|^2 + |\overline{Y}_p|^2} \Bigg) \end{aligned} \tag{3}$$

In (3) each label i is represented by two output neurons (with indexes $2i$ and $2i+1$). The first of them represents the "degree of assignment" of the ith label to a given input, and the other is a threshold associated with this label. The assignment of the set of labels to a given test input x has changed accordingly from (2) to $h(x) = \{q \in Y : c_{2q}(x) > c_{2q+1}(x)\}$. Furthermore, as opposed to (1), all possible pairs of labels are considered in (3).

The above formulation was revisited and further tuned in [20], leading to the Tuned CART-M method (TCART-M), by introduction of a scaling parameter D which maintains a balance between the core error formulation (1) and the components added in (3):

$$E_{TCART-M} = \sum_{p=1}^{m} \Bigg(\frac{\sum\limits_{(r,s)\in Y_p \times \overline{Y}_p} (e^{-(c^p_{2r}-c^p_{2s})} + \frac{e^{-(c^p_{2s+1}-c^p_{2r+1})}}{D})}{2|Y_p||\overline{Y}_p| + |Y_p|^2 + |\overline{Y}_p|^2} + \frac{\sum\limits_{r\in Y_p}\sum\limits_{t\in Y_p} \frac{e^{-(c^p_{2r}-c^p_{2t+1})}}{D} + \sum\limits_{s\in \overline{Y}_p}\sum\limits_{t\in \overline{Y}_p} \frac{e^{-(c^p_{2t+1}-c^p_{2s})}}{D}}{2|Y_p||\overline{Y}_p| + |Y_p|^2 + |\overline{Y}_p|^2} \Bigg) \tag{4}$$

Parameter D is autonomously fine-tuned by the system for a given data set in the nested cross-validation process. Two versions of the tuning method are presented in [20]: TCART-M*g* in which D is optimized independently of the choice of an evaluation measure, and TCART-M*i* which optimizes D for a particular error measure. Due to higher generality and practical relevance of the former approach, in the rest of this paper the TCART-M*g* version is considered and referred to as TCART-M.

A. TCART-M in dimensionality reduction (TCART-MR)

The main goal of this paper is extension of the TCART-M method to DR task. Application of TCART-M to DR is considered in the two following contexts:

- as an auxiliary preprocessing phase of the TCART-MR usage in MC, or
- as a goal *per se* leading to higher compactness of data representation. In this case, the DR phase of TCART-MR can be subsequently followed by application of *any* MC method.

In the proposed method the importance of particular input features in the trained MLP network is estimated in a straightforward way based on inspection of the weights coming out from the respective input neurons. More precisely, for each input $i = 1, \ldots, n$ we define its *utility* value u_i in the following way: $u_i = \sum_{j=1}^{m} |w^1_{ij}|$, where m is the size of the hidden layer and w^1_{ij} is connection weight between the ith input neuron and the jth 1hl unit. The *utility* value is used for selection of the most pertinent features from the input data. The greater the u_i value, the more relevant the ith input is assumed to be.

Once *utility* values for all inputs are calculated the last decision is the choice of the target input dimensionality $\rho < n$. To this end we adopt the *golden-section search algorithm* [36] which recursively narrows the range of values inside which the extremum is known to exist. The usage of the method is further discussed in Section V-A. It is proven that the number of iterations required to find the optimum of unimodal function on $[a, b]$ interval with ε accuracy using the golden-section search is equal to $\left\lceil \log_k \frac{\varepsilon}{b-a} \right\rceil$, $k \approx 0.618$. In our case $\varepsilon = 1$ and the size of the interval being searched is equal to n (the size of the input), what leads to $\lceil \log_k \frac{1}{n} \rceil$. In order to get the above mentioned function value, in each iteration TCART-M method has to be run. Thus, the number of TCART-M calls is logarithmic. Please observe, that if the assumption about unimodality does not hold, the algorithm still finds a local optimum, which should anyway lead to improvement of results compared to the case without dimensionality reduction.

The above procedure of detection of the subset of potentially most relevant inputs (together with automatic calculation of the size of this subset of input features) is applied to the trained TCART-M network and once this subset is selected the TCART-M training method is applied again on the input data with reduced dimensionality. This *nested* application of the TCART-M training method constitutes a new MC approach, referred to as TCART-MR in the experimental evaluation sections.

B. TCART-M and TCART-MR in multilabel classification

In [20], TCART-M method was compared with 12 most popular MC approaches based on 5 widely-used benchmark problems and an ensemble of 16 error measures. The experimental setup and detailed results can be retrieved from the source paper. On a general note, in a cumulative assessment comparison received by summing up ranking positions of each MC method over all benchmark sets and all error measures, TCART-M appeared to be the best approach, followed by CLR [23] and BR [3].

IV. Experimental setup

A. Benchmark problems

Nine popular MC benchmark problems, summarized in Table I, were selected for experimental evaluation of the TCART-MR method versus state-of-the-art MC approaches. The data sets come from various domains and differ in the number of labels, training samples, attributes, as well as the average number of labels per class.

Name	**Domain**	N	n	Q	**Avg. Card.**	**Ref.**
emotions	audio	593	72	6	1.87	[1]
yeast	biology	2417	103	14	4.24	[8]
scene	images	2407	294	6	1.07	[3]
enron	text	1702	1001	53	3.38	[6]
medical	text	978	1449	45	1.25	[7]
flags	images	194	19	7	3.392	[37]
birds	audio	645	260	19	1.014	[38]
genbase	biology	662	1186	27	1.252	[39]
CAL500	text	502	68	174	26.044	[40]

TABLE I: Basic parameters of the benchmark data sets: name, domain of origin, number of instances, number of attributes, number of labels, and the average number of labels per class.

B. Evaluation measures

In traditional binary or multiclass classification there are certain well-established metrics, e.g. *accuracy* or *loss*, which are commonly applied to evaluate classification quality. Due to higher practical complexity of MC, stemming from the necessity of assignment of variable number of labels to particular data instances, both the number and variety of MC-related baseline metrics are higher.

In order to make the evaluation process fair and comprehensive, we followed the approach proposed in the recent MC survey [21] in which 16 variable error measures were applied allowing for examination of the tested methods from various perspectives. These metrics are grouped into three categories and the results are reported on the three levels of detail: in the form of one cumulative score, which combines results of all 16 error measures, as three group-based indicators (one per each group of error measures), and on a detailed level - separately for each metrics. An in-depth description of all 16 metrics can be found in [21]. Values of all measures except the *coverage* fit the interval $[0, 1]$. For *Hamming loss*, *one-error*, *coverage* and *ranking loss*, the smaller the value, the better the assessed method's quality. In the remaining cases the greater the value, the higher the estimated performance.

V. Experimental evaluation

A. Experimental results in dimensionality reduction

In order to evaluate DR efficacy of TCART-MR, it was compared with three state-of-the-art DR methods: PCA - which is the most popular, general purpose DR approach, MLSI [17] and MDDM [16] - which are, in turn, the most popular DR algorithms among those designed specifically for the MC domain. MLSI is a supervised extension of LSI which retains statistical information not only about the input features but also about the multivariate outputs. A mapping of the input features onto a new feature space with lower dimensionality is derived by solving a linear optimization problem.

MDDM finds a lower-dimensional input space by maximizing a dependence (averaged over all examples) between feature space and associated labels. Finding a reduced space is performed by solving eigen-decomposition problem.

In the case of PCA, after input space transformation, principal components (PCs) with the lowest variance were removed and all the remaining ones served as the new input data for the TCART-M algorithm. In order to make the comparison fair, the number of selected PCs was chosen in a greedy manner, i.e. the number of them that led to the best results were selected. The same exhaustive strategy was applied to MLSI and MDDM, i.e. the input dimensionality that led to the best outcome was selected for comparison with TCART-MR. MLSI parameters were set according to [17]: $\beta = 0.5$, $\gamma = 0$. MDDM was implemented with uncorrelated projection constraint denoted by MDDM$_p$ in [16].

Recall that the method of dimensionality reduction of the MC data relies on the *utility* values $u_i, i = 1, \ldots, n$ assigned to MC input features and calculated based on the respective weight values of the neural network trained with TCART-M algorithm (see Section III-A). The reduced input representation is composed of the features corresponding to the top ρ *utility* values. The process of finding the optimal number of input features for a given benchmark is automatized in this paper (TCART-MR approach) by adopting the *golden-section search algorithm*, as mentioned in Section III-A. Figure 1 presents a relation between a degree of DR and the TCART-M performance, for 4 example benchmarks. The optimal number of reduced features clearly depends on a particular data set, but to a much lesser extent on a particular error measure.

Optimal numbers of removed input features for all 9 benchmarks and all four tested methods are presented in Table II. The values range from 21%, for *flags* data set, up to 97%, for *medical* data set, whose samples contain over 1400 features, representing the numbers of all word occurrences in the patient's symptoms description (including single digits, punctuation marks, etc.). It stems from Table II that there is no straightforward relation between data set properties (numbers of attributes or labels, average label cardinality, or domain of the data) and the optimal number of removed features. Estimation of the optimal number of features depends on nonlinear inter-dependencies among features reflecting the nature (inner structure) of the data and its domain of origin.

	TCART-MR	**PCA**	**MLSI**	**MDDM**
emotions	25 (35%)	36 (50%)	34 (47%)	31 (43%)
yeast	41 (40%)	67 (65%)	50 (48%)	45 (44%)
scene	147 (50%)	250 (85%)	199 (68%)	153 (52%)
enron	650 (65%)	720 (72%)	705 (70%)	620 (62%)
medical	1376 (95%)	1405 (97%)	1357 (94%)	1333 (92%)
flags	4 (21%)	5 (28%)	4 (21%)	5 (28%)
birds	143 (55%)	166 (64%)	145 (56%)	161 (62%)
genbase	498 (42%)	534 (45%)	489 (41%)	522 (44%)
CAL500	30 (44%)	35 (52%)	34 (50%)	32 (47%)

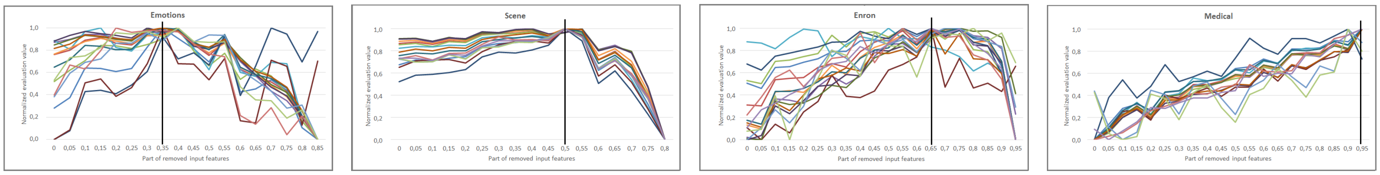
TABLE II: Optimal numbers of removed input features in TCART-MR method and in other DR algorithms used in the input preprocessing phase before a TCART-M classification.

In Table III a *cumulative DR score* across all benchmarks is calculated as a sum of ranking positions of a given method for all 16 evaluation measures. TCART-MR turned out to be the best method for 7 benchmarks, and in 3 of them (*emotions*, *scene* and *enron*) gained the 1st position for each evaluation measure. Based on 1-tailed t-test with significance level equal to 0.05 for 5 data sets (*emotions*, *scene*, *enron*, *medical* and *flags*), out of these 7, the results are statistically significant. Normal distribution of results (which is the requirement for using t-test prerequisite) was checked by Shapiro-Wilk test.

The advantage of TCART-MR over PCA stems from taking into account the input-output relation between data samples and label-sets during the training process, as opposed to PCA, which is an unsupervised method independent of the output labels. The superiority of TCART-MR over the two other DR methods is attributed to its underlying non-linear nature, contrary to MLSI which relies on linear correlations only. Furthermore, the ability of TCART-MR to learn the input-output relation *directly from data samples* seems to be a decisive factor in its overall superiority. Finally, both TCART-MR training phases (the initial DR phase and the final MC one) are performed with the same neural architecture (differing by the input size only) what makes the whole process more coherent than in the competitive cases, where DR and MC phases are performed using different methods.

B. Experimental results in multilabel classification

This section extends our previous experiments presented in [20] by making comparison of the proposed TCART-MR method with 11 MC approaches - three neural network based (BP-MLL, CART-M, TCART-M) described in Section III and eight other, state-of-the-art methods presented in [21], briefly introduced in Section II, whose implementations were obtained



(a) Emotions

(b) Scene

(c) Enron

(d) Medical

Fig. 1: Relationship between TCART-M quality (for each of 16 error measures) and the fraction of removed input features for 4 selected benchmarks. All evaluation measures are normalized to [0,1] interval.

	TCART-M	TCART-MR	PCA	MLSI	MDDM
emotions	46	**16**	80	54	41
yeast	63	47	43	48	**34**
scene	72	**16**	64	45	40
enron	77	**16**	61	43	38
medical	80	**20**	57	44	39
flags	64	**22**	58	48	46
birds	68	**23**	54	50	45
genbase	78	34	48	50	**30**
CAL500	80	**36**	50	**36**	38
sum	628	**230**	515	418	351

TABLE III: Comparison of DR methods by summing their ranking positions for all evaluation measures. The value of each score is between 16 (always the 1st place) and 80 (always the last one). For each benchmark, the best method is **bolded**. Gray background denotes statistical significance of the result difference between TCART-MR and a given method according to 1-tailed t-test with significance level equal to 0.05.

from *Mulan* library [41]. The results are presented in an aggregated way: for each benchmark set (in the form of cumulative scores with respect to all error measures), and for each error metric (as cumulative scores with respect to all benchmark sets). For detailed outcomes please visit [42].

Table IV, for each tested method and each benchmark set, presents sums of ranking positions across all 16 evaluation measures. For each benchmark and each error measure, the respective ranking was created based on the average values of the considered error measure from 30 independent runs. Afterwards these 16 ranking positions were summed up to yield a cumulative score. In this comparison, in case of 3 out of 9 benchmarks TCART-MR gained the first place and its overall score across all 9 benchmarks is the lowest (the best one) with more than 100 points ahead of the runner-up method (CLR). TCART-MR is also clearly the most effective among neural network MC approaches (the first 4 leftmost methods). The baseline version TCART-M, gained the 5th overall position.

The same experimental data is presented in Table V, though from a different perspective. Each position in the table presents the sum of ranking positions across all 9 benchmark sets, for the respective method and evaluation metric. Since the table refers to the same data as Table IV, but structured differently, for each method its overall cumulative score (the last row) does not change, and obviously TCART-MR remains the best-scoring approach. In individual inspection of particular evaluation measures TCART-MR gained the 1st place in 3 out of 16 error measures and was excelled in this comparison only by the HOMER [28] approach with 5 winning positions.

One of the general conclusions from presented results is the importance of thorough evaluation of MC approaches in terms of diversity of benchmark data and error measures. Analysis of results presented in both tables indicates a complementary role of variable benchmark selection, on the one hand, and a wide range of specific evaluation metrics, on the other hand. As can be easily concluded from the tables none of the algorithms is superior over the others in a wide range of data sets or evaluation measures. These observations are confirmed by the Friedman non-parametric statistical test with Nemenyi post-hoc analysis which proved that none of the methods statistically significantly differ from all the others (test confirmed the null hypothesis that methods performances are similar with $\alpha = 0.05$). On the other hand, despite the lack of statistical superiority of any method over the others across all benchmarks and all error measures, certain differences between methods can clearly be observed and supported by experimental evidence. For instance, in terms of benchmark sets comparison (Table IV), TCART-MR was the only method which won against the others in 3 cases, ECC [22] won twice, and each of the remaining approaches at most once.

Strong dependence between the method's performance and the type of the error function is also a straightforward observation. In the class of *example-based* measures (top 6 rows) the leading method is ECC with 223 points, for *label-based* measures (next 6 rows) it is TCART-MR with the score of 270, and for *ranking-based* measures (last 4 rows) the best approach is CLR which scored 137 points. Overall, the winning method - TCART-MR - seems to be the most universal or least prone to selection of the error measure as it accomplished the 3rd, 1st and 2nd places, resp. in *example-based*, *label-based* and *ranking-based* measures, with only minute score differences compared to the leading methods.

While the above presented results and their analysis confirm that TCART-MR is the most universal approach among the tested ones, it should be admitted that the method is generally less efficient for the data sets with large numbers of input attributes. In such cases we recommend to use HOMER [28] approach, which was designed specifically for highly-dimensional MC data sets and additionally optimized

	BP-MLL	CART-M	TCART-M	TCART-MR	BR	CC	CLR	HOMER	ML-k NN	RAkEL	RF-PCT	ECC
emotions	64	70	41	**25**	130	139	134	125	181	133	65	139
yeast	117	98	114	107	83	95	**62**	92	126	112	125	112
scene	159	117	104	79	77	75	93	95	161	**70**	135	80
enron	136	97	83	**51**	82	122	83	96	157	127	100	110
medical	153	133	124	99	108	96	69	**62**	117	89	85	103
flags	161	148	109	**44**	106	133	56	120	80	117	92	70
birds	192	139	117	88	82	114	71	101	106	93	91	**42**
genbase	175	133	96	90	80	38	96	115	164	46	114	**31**
CAL500	**60**	124	102	71	98	94	112	101	109	104	100	96
sum	1217	1059	890	**654**	846	906	776	907	1201	891	907	783

TABLE IV: Benchmark-based comparison of neural network approaches (columns 1-4) with the state-of-the art MC approaches. Each value denotes the sum of ranking positions across 16 evaluation measures for given benchmark and MC method. Best values for each benchmark are **bolded**. Gray background denotes statistical significance of the results difference between TCART-MR and a given method according to 1-tailed t-test with significance level equal to 0.05.

	BP-MLL	CART-M	TCART-M	TCART-MR	BR	CC	CLR	HOMER	ML-k NN	RAkEL	RF-PCT	ECC
Hamming Loss	100	58	54	41	48	67	**40**	75	69	**40**	46	47
Accuracy	78	78	62	38	62	59	54	29	85	42	80	**27**
Precision	98	68	47	**34**	53	63	48	74	69	39	60	42
Recall	38	84	66	51	59	56	53	**24**	89	47	86	42
Subset Accuracy	84	61	55	37	49	50	56	40	70	**28**	63	27
F1 score	85	78	59	39	56	59	53	**29**	79	39	81	38
Micro-precision	104	46	54	35	59	66	58	82	51	50	**26**	57
Macro-precision	90	63	58	51	33	42	51	77	80	58	**26**	58
Micro-recall	43	80	65	54	57	55	53	**24**	93	47	80	43
Macro-recall	42	63	55	44	57	52	62	**33**	100	55	81	52
Micro-F1	78	81	66	46	56	64	51	**31**	88	40	59	35
Macro-F1	63	64	52	**40**	54	48	60	41	98	57	58	58
Ranking Loss	77	62	48	38	52	52	**35**	87	56	90	38	66
OneError	81	56	49	33	53	63	**32**	86	59	79	47	60
Coverage	74	59	52	43	55	57	39	88	51	90	**27**	66
Average Precision	82	58	48	**30**	43	53	31	87	64	90	49	65
sum	1217	1059	890	**654**	846	906	776	907	1201	891	907	783

TABLE V: Error measure-based comparison of neural network approaches (columns 1-4) with the state-of-the art MC approaches. Each value denotes the sum of ranking positions across 9 benchmark sets for given evaluation measure and MC method. Best values for each error measure are **bolded**. Gray background denotes statistical significance of the results difference between TCART-MR and a given method according to 1-tailed t-test with significance level equal to 0.05.

for computational efficiency.

VI. Conclusions

This paper introduces and evaluates a new neural network method (TCART-MR) for dimensionality reduction of the input feature space in MC domain. The proposed approach extends our previous MC algorithm TCART-M [20] by adding a procedure of weight examination of the trained TCART-M model so as to select the optimal number of the most relevant input features. Once the subset of features is selected the training is repeated on the data with reduced input dimensionality.

TCART-MR was tested on a set of 9 well-established benchmark problems along two directions: quality of DR process and the overall efficacy in MC domain. With respect to DR, experimental results showed statistically relevant advantage of TCART-MR over three competitive, widely-known DR approaches: PCA - the most commonly used general purpose DR method, MLSI and MDDM - the state-of-the-art input reduction methods specifically developed for the MC task. In terms of MC quality, TCART-MR not only outperformed its predecessors (BP-MLL, CART-M and TCART-M), but also achieved the best result in comparison with 8 other, state-of-the-art MC approaches listed in the recent MC survey paper [21]. Gaining the first place in the *cumulative score*, which relies on ranking positions of a particular method for all 9 benchmarks and against 16 variable evaluation measures, confirmed universality and flexibility of the proposed method.

On a general note, the results suggest that a universal MC approach suitable for a wide range of data sets and error measures is unlikely to exist. Consequently, testing new methods on a wide range of benchmarks and against various error metrics is of particular value in this domain.

Another general conclusion refers to the specificity of benchmark sets used in MC domain. In majority of the tested cases removing over 40% of the least relevant (according to TCART-MR) input features did not cause deterioration of results and usually even led to their improvement. The reason of this phenomenon is most probably the real-life origin of the test data, which is not filtered in any specific way. In this context the DR preprocessing step is of paramount importance.

Our current focus is on alternative, non-gradient optimization methods (e.g. genetic algorithms, simulated annealing or tabu search) for setting dimensionality threshold (ρ) in the TCART-MR method, in place of the *golden-section search algorithm*.

Acknowledgment

This work was supported by the National Science Centre, Poland, grant number 2017/25/B/ST6/02061.

References

[1] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," in *Proceedings of the 9th International Conference on Music Information Retrieval*, Philadelphia, USA, September 14-18 2008, pp. 325–330.

[2] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM International Conference on Multimedia*, 2006, pp. 421–430.

[3] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[4] A. N. Srivastava and B. Zane-Ulman, "Discovering recurring anomalies in text reports regarding complex space systems," in *IEEE Aerospace Conference.*, 2005, p. 55–63.

[5] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proceedings of the ECML/PKDD 2008 Discovery Challenge*, 2008, p. 75–84.

[6] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Prof. 15th European Conference on Machine Learning, Pisa, Italy*. Springer Berlin Heidelberg, 2004, pp. 217–226.

[7] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch, "A shared task involving multi-label classification of clinical free text," in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, ser. BioNLP '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 97–104.

[8] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *In Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 681–687.

[9] I. T. Jolliffe, *Principal component analysis*. Springer, 1986.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[12] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, "Dimensionality reduction using non-negative matrix factorization for information retrieval," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 960–965.

[13] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2010, pp. 126–139.

[14] Y.-N. Chen and H.-T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 1529–1537.

[15] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 381–389.

[16] Y. Zhang and Z.-H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 3, p. 14, 2010.

[17] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 258–265.

[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[19] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.

[20] J. Mańdziuk, A. Żychowski, and L. Wang, "A TCART-M - Tuned CARTesian-based Error Function for Multilabel Classification with the MLP," in *International Joint Conference on Neural Networks (IJCNN'2017)*, vol. 5199. Anchorage, AK, USA: IEEE Press, 2017, pp. 565–572.

[21] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084 – 3104, 2012.

[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. ECML/PKDD, Bled, Slovenia*. Springer Berlin Heidelberg, 2009, pp. 254–269.

[23] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.

[24] S.-H. Park and J. Fürnkranz, "Efficient pairwise classification," in *European Conference on Machine Learning*. Springer, 2007, pp. 658–665.

[25] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038 – 2048, 2007.

[26] ——, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[27] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *European Conference on Machine Learning*. Springer, 2007, pp. 406–417.

[28] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, 2008, pp. 30–44.

[29] H. Blockeel, L. De Raedt, and J. Ramon, "Top-down induction of clustering trees," *arXiv preprint cs/0011032*, 2000.

[30] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 115–124.

[31] Y. Yan, Y. Wang, W.-C. Gao, B.-W. Zhang, C. Yang, and X.-C. Yin, "LSTM^2: Multi-label ranking for document classification," *Neural Processing Letters*, vol. 47, no. 1, pp. 117–138, 2018.

[32] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-Ray classification," *arXiv preprint arXiv:1803.02315*, 2018.

[33] A. Maxwell, R. Li, B. Yang, H. Weng, A. Ou, H. Hong, Z. Zhou, P. Gong, and C. Zhang, "Deep learning architectures for multi-label classification of intelligent health risk prediction," *BMC bioinformatics*, vol. 18, no. 14, p. 523, 2017.

[34] Y. Gong, K. H. T. Leung, A. T. Toshev, S. Ioffe, and Y. Jia, "Ranking approach to train deep neural nets for multilabel image annotation," Jan. 24 2017, uS Patent 9,552,549.

[35] R. Grodzicki, J. Mańdziuk, and L. Wang, "Improved Multilabel Classification with Neural Networks," in *Parallel Problem Solving from Nature*, ser. Lecture Notes in Computer Science, vol. 5199. Springer Verlag, 2008, pp. 409–416.

[36] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical recipes," 1989.

[37] E. C. Gonçalves, A. Plastino, and A. A. Freitas, "A genetic algorithm for optimizing the label ordering in multi-label classifier chains," in *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*. IEEE, 2013, pp. 469–476.

[38] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern *et al.*, "The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *Machine Learning*

for Signal Processing (MLSP), 2013 IEEE International Workshop on. IEEE, 2013, pp. 1–8.

[39] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms," in *Panhellenic Conference on Informatics*. Springer, 2005, pp. 448–456.

[40] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

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

[42] "Multilabel Classification with neural networks. Detailed experimental results of TCART-M and TCART-MR methods," Accessed: 2018-12-03. [Online]. Available: http://www.mini.pw.edu.pl/~mandziuk/multilabel/detailed_results.pdf