ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Efficient feature selection for domain adaptation using Mutual Information Maximization

Guillermo Castillo-García, Laura Morán-Fernández, Verónica Bolón-Canedo *

CITIC, Universidade da Coruña, A Coruña, Spain

Abstract. Green AI, an emerging research field, focuses on improving the efficiency of machine learning models. In this paper, we introduce a novel and efficient method for feature selection in domain adaptation, a type of transfer learning where the source and target domains share the feature space and task but differ in their distributions. Instead of using evolutionary algorithms, a typical approach in this field, we propose the use of filter methods, which do not require an iterative search process and are less computationally expensive. Our proposed method is Mutual Information Maximization, and our experiments show that it outperforms Particle Swarm Optimization in terms of efficiency, speed, and the ability to select a reduced subset of features while achieving competitive classification accuracy results.

1 Introduction

With the increasing energy demands of modern machine learning techniques, researchers are exploring ways to improve the efficiency of these methods. This approach is called Green AI [1], and includes optimizing hardware, reducing the size of models, or developing new algorithms that require less computational power. By prioritizing energy efficiency in machine learning not only we can reduce its carbon footprint, but also increase its accessibility and affordability.

Feature selection [2] is a technique used in machine learning to reduce the dimensionality of a dataset, with the goal of selecting the features that provide useful information for our predictive model, and therefore reducing the amount of data used. On the other hand, transfer learning is another method that aims to make use of already learned knowledge for one domain in a different domain. The case in which we encounter a shared task and feature space in both the source and target datasets, but their distributions differ, is referred to as domain adaptation. The objective of feature-based transfer learning approaches is to identify a common feature representation that minimizes the distributional discrepancy between the source and target data while preserving the important predictive learning abilities in both datasets. Our aim is to apply feature selection techniques to find this common feature representation.

Highly dimensional feature selection problems with complex feature interactions have been widely addressed using Evolutionary Computation techniques.

*This work was supported by the Ministry of Science and Innovation of Spain (Grant PID2019-109238GB-C22 / AEI / 10.13039 / 501100011033) and together with "NextGenerationE"/PRTR (TED2021-130599A-I00) and by Xunta de Galicia (Grants ED431G 2019/01 and ED431C 2022/44).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

However, this kind of algorithms can be computationally expensive due to their iterative search process. Examples of these methods could be Differential Evolution [3] or Particle Swarm Optimization [4]. Since our goal is to improve the efficiency of the domain adaptation task, we propose to use Mutual Information Maximization, a filter method that does not require iterative searches and is therefore significantly less computationally demanding. In this paper, we therefore compare the performance and efficiency of an evolutionary approach, Particle Swarm Optimization, and Mutual Information Maximization [5] for performing feature selection for domain adaptation.

2 Feature-based domain adaptation

Transfer learning refers to the procedure of utilizing the relevant information from a source domain D^S and source task T^S in order to enhance the performance of a target predictive function $f^T(\cdot)$, given a target domain D^T and task T^T, where $D^S \neq D^T$ or $T^S \neq T^T$.

The objective of feature-based transfer learning is to identify a feature representation that can achieve substantial predictive accuracy in both the source and target domains, while also minimizing the differences between their data distributions. For the case of domain adaptation, the techniques used aim to extract common and informative features from both domains, therefore reducing the differences between their marginal distributions. The overall process we used for performing feature-based domain adaptation is depicted in Figure 1.

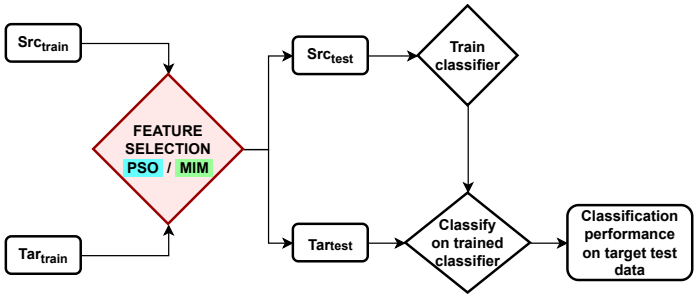

Fig. 1: Diagram of the process of feature selection for domain adaptation

As we can see, we start by performing feature selection using our source and target training data, for which we are going to compare two approaches, using Particle Swarm Optimization or Mutual Information Maximization, each following a different process, as we will see below. This results in a subset of features that will be the one used in our testing data, both on the source and target domains. We then train a classifier with the source test data, and classify the target test data, giving us the performance on the test phase.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) [4] is an algorithm for solving optimization problems inspired by the behaviour of a flock of birds. It sets a swarm of particles which explore the search space in parallel. Each particle is a solution candidate, and consists of a position and velocity (or momentum). The value of each particle is computed using a fitness function, which guides them to a good solution. It was originally proposed to solve continuous problems, therefore we opted to use Sticky Binary Particle Swarm Optimization (SBPSO) [6], which redefines the concept of momentum making it appropriate for binary problems.

For this study, we used the approach presented in a previous work [7], which includes using classification accuracy or data complexity metrics in the fitness function, which is the following:

$$Fitness = sw * srcErr + tw * tarErr + stw * diffST \tag{1}$$

where sw, tw and stw are weights, $srcErr$ and $tarErr$ are classification errors on source and target data (or complexity measures with lower values indicating less complexity), and $diffST$ measures the difference in the marginal distributions of each data partition using Maximum Mean Discrepancy with Gaussian Radial Basis function as the kernel function. The difference between the two SBPSO versions used in this study is that one uses classifiers to calculate $srcErr$ and $tarErr$, and the other uses complexity measures instead.

We will use the best performing option for each case according to [7], that is, a k-Nearest Neighbors (kNN) classifier, which uses proximity between data points to predict the class of another, and the Maximum individual feature efficiency (F3) complexity measure, which estimates the individual efficiency of each feature in separating the classes.

2.2 Mutual Information Maximization

Mutual Information (MI) is a measure of the mutual dependence between two features, one of them being the class to be predicted in our case. Given a feature X and the class label Y, with probability mass functions $p(x)$ and $p(y)$, respectively, and joint probability mass function $p(x, y)$, the mutual information between X and Y ($I(X;Y)$) is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

Mutual Information Maximization (MIM) ranks all the features based on their MI score and selects the top k features, $MIM(X_k) = I(X_k;Y)$. The value of k is determined either by a predefined requirement for a specific number of features or by some other stopping criterion. A limitation of this method is that it assumes that each feature is independent of all the others, thus not taking into account the redundancy between them.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

To perform feature selection for domain adaptation using MIM, we first calculate MIM for the source dataset (Src), and select the top 10% features in the resulting ranking. Then, we do the same for the target dataset (Tar), also selecting the top 10% features. Then, the resulting subset of selected features is the union of both subsets, which would mean selecting up to 20% of all features, although this will rarely happen (it did not in our experiments), since it is to be expected (and is ultimately the goal) to have common features selected from both Src and Tar domains.

3 Experimental design

We tested the proposed approaches in two different problems. The first one is handwritten digits recognition, for which we used two datasets: MNIST [8] and USPS [9]. Both of them are composed of samples with 256 features (16x16 pixels) and 10 possible classes, one for each digit. Their distributions differ, as they are collected from different sources. We used 800 samples of MNIST, and 720 of USPS, using one dataset as source data and the other as target. The second problem is predicting if a restaurant will be liked based only on its characteristics and price range. For that, we used data from TripAdvisor [10], which contains restaurant reviews from different cities, with 209 features and two possible classes. We used the datasets corresponding to Madrid (Mad) and Barcelona (Bar), using a subsample of 1000 instances for each city and, as in the previous case, using one of them as source and the other as target data. In both problems, 70% of the data was used for training and 30% for testing.

With regard to the experimental settings, first, a process to fine tune the parameters was carried out to select the best hyperparameters for each method. Then, for each problem, we performed 15 tests with the same 15 random seeds for each method, following the process presented in Section 2. For doing feature selection, we employ the three methods mentioned previously: SBPSO using kNN, SBPSO using F3, and MIM. To evaluate the quality of the selected features we will use the classification accuracy. Therefore, we will use the classifiers k-Nearest Neighbor (kNN), Support Vector Machines (SVM) and Naive-Bayes (NB), in order to get a broader view of how the models perform and to detect the ones that are dependent on the classifier used.

4 Experiments

The results obtained in the testing phase are shown in Table 1. We can see that, using MNIST as source dataset and USPS as target, SBPSO-kNN performs better when using kNN as subsequent classifier. This was expected, as using a classifier for training can provide an advantage when testing on that same classifier. When classifying with SVM and NB, MIM obtains better results than both SBPSO options. Despite that, SBPSO-kNN achieves the best results when using USPS as source and MNIST as target, but MIM is close, performing better than SBPSO-F3 with the three classifiers used for testing.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Table 1: Mean classification accuracy, percentage of features selected and time elapsed (seconds). The best results for each dataset are marked in bold.

Src-Tar	Method	kNN	SVM	NB	% Feat.	Time
MNIST-USPS	SBPSO-kNN	**0.600**	0.433	0.528	51.76	555.75
MNIST-USPS	SBPSO-F3	0.533	0.367	0.490	32.68	356.00
MNIST-USPS	MIM	0.546	**0.531**	**0.530**	**18.20**	**2.88**
USPS-MNIST	SBPSO-kNN	**0.383**	**0.402**	**0.351**	45.93	1050.06
USPS-MNIST	SBPSO-F3	0.299	0.300	0.326	27.99	347.88
USPS-MNIST	MIM	0.366	0.376	0.335	**18.49**	**2.87**
Mad-Bar	SBPSO-kNN	**0.996**	**0.994**	0.875	64.33	1800.26
Mad-Bar	SBPSO-F3	0.992	0.992	**0.985**	34.07	642.32
Mad-Bar	MIM	0.938	0.956	0.965	**18.72**	**1.13**
Bar-Mad	SBPSO-kNN	**0.991**	**0.990**	0.879	67.07	1735.44
Bar-Mad	SBPSO-F3	0.988	0.989	**0.976**	33.13	698.25
Bar-Mad	MIM	0.923	0.955	0.950	**18.72**	**1.12**

On the TripAdvisor problem, SBPSO-kNN obtains the best results when using kNN and SVM as classifiers, but shows the lowest accuracy with NB. In both cases of this problem (Bar-Mad and Mad-Bar), MIM performs slightly worse than SBPSO-F3, but the results are competitive, with accuracy over 92% in all cases.

Focusing on efficiency, MIM is clearly the best option. It always selects the lowest number of features while maintaining a competitive performance, and provides a great advantage in terms of training time needed. In all cases, when compared to both SBPSO options, MIM results in a time reduction between 99.19% and 99.94%. Considering the potential loss in accuracy (if there is any, as we have seen that it can outperform SBPSO in some cases) there was, in the worst case (SBPSO-kNN with Bar-Mad datasets), a reduction of 6.8%. These findings suggest that the benefits gained in terms of efficiency outweigh the potential loss in classification performance.

5 Conclusions

As machine learning models get bigger and bigger, the need for Green AI grows, leading to research on improving the efficiency of these methods. In this paper, we focused on the efficiency of domain adaptation algorithms, making a comparison between different methods. Due to the high computational cost of state of the art methods such as evolutionary algorithms, we suggest a more efficient approach based on filter methods for feature selection, utilizing MIM. In this study, we conducted a comparative analysis between MIM and two previously proposed alternatives based on an evolutionary algorithm: Sticky Binary Particle Swarm Optimization using classifiers (kNN) or data complexity metrics (F3) in the fitness function.

After carrying out experiments over two datasets (leading to four different

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

scenarios), we proved that MIM is competitive in terms of accuracy, even performing better than SBPSO with F3 on the Handwritten Digits problem, and is able to dramatically reduce the time required for training and the number of features selected. Moreover, when compared to the version of SBPSO that uses a classifier in the fitness function, our proposed method has the advantage of being independent of the classifier used in a posterior test phase.

Based on these findings, we conclude that our proposed approach provides a considerably more efficient alternative to evolutionary algorithms in domain adaptation. Additionally, the slight decrease in accuracy observed is outweighed by the significant gains in time efficiency. Therefore, depending on the requirements of the task at hand, this trade-off may be highly recommended.

References

[1] R. Schwartz et al. "Green AI". In: *Communications of the ACM* 63.12 (2020), pp. 54–63.

[2] I. Guyon et al. *Feature extraction: foundations and applications*. Vol. 207. Springer, 2008.

[3] R. Storn and K. Price. "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces". In: *Journal of global optimization* 11.4 (1997), p. 341.

[4] J. Kennedy and R. Eberhart. "Particle swarm optimization". In: *Proc. of ICNN'95 Int. Conf. on Neural Networks*. Vol. 4. IEEE. 1995, pp. 1942–1948.

[5] J. Vergara and P. Estévez. "A review of feature selection methods based on mutual information". In: *Neural computing and applications* 24 (2014), pp. 175–186.

[6] B.H. Nguyen, B. Xue, and P. Andreae. "A particle swarm optimization based feature selection approach to transfer learning in classification". In: *Proc. of Genetic and Evolutionary Computation Conf.* 2018, pp. 37–44.

[7] G. Castillo-García, L. Morán-Fernández, and V. Bolón-Canedo. "Feature selection for transfer learning using particle swarm optimization and complexity measures". In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2022, pp. 7–12.

[8] L. Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[9] J. J. Hull. "A database for handwritten text recognition research". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.5 (1994), pp. 550–554. DOI: 10.1109/34.291440.

[10] P. Pérez-Núñez et al. *TripAdvisor Restaurant Reviews*. Version 1.0.0. Zenodo, Dec. 2021. DOI: 10.5281/zenodo.5644892. URL: https://doi.org/10.5281/zenodo.5644892.