Knowledge-Based Systems 000 (2018) 1–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Statistical comparisons of active learning strategies over multiple datasets

Oscar Reyes[a,c], Abdulrahman H. Altalhi[b], Sebastián Ventura[a,b,c,*]

[a] Department of Computer Science and Numerical Analysis, University of Córdoba, Spain
[b] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[c] Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimonides Biomedical Research Institute of Córdoba, Spain

ARTICLE INFO

Article history:
Received 21 September 2017
Revised 26 January 2018
Accepted 29 January 2018
Available online xxx

Keywords:
Active learning
Statistical comparison
Non-parametric statistical tests

ABSTRACT

Active learning has become an important area of research owing to the increasing number of real-world problems in which a huge amount of unlabelled data is available. Active learning strategies are commonly compared by means of visually comparing learning curves. However, in cases where several active learning strategies are assessed on multiple datasets, the visual comparison of learning curves may not be the best choice to conclude whether a strategy is significantly better than another one. In this paper, two comparison approaches are proposed, based on the use of non-parametric statistical tests, to statistically compare active learning strategies over multiple datasets. The application of the two approaches is illustrated by means of a thorough experimental study, demonstrating the usefulness of the proposal for the analysis of the active learning performance.

1. Introduction

Machine learning aims at constructing computational algorithms capable of determining general patterns from available data. In the learning process, not all data are useful, because noisy, redundant and incomplete data can affect in many ways the performance of a learning algorithm. Consequently, the acquisition of a high-quality and compact dataset (a.k.a. training set) from which the learning algorithm can determine useful patterns is of the most importance [70,75].

Sample selection is a crucial preprocessing step in data mining [66]. Sample selection methods aim at selecting select a representative subset from the original dataset, in such a manner that the performance of the learner generated from the selected subset would be similar (or even better) than on the original dataset. The main advantages in applying sample selection methods are as follows [39,49,70,76]: (I) reduction of the storage requirements by removing redundant information from datasets, (II) decrease in computation efforts for predicting new patterns, (III) improvement of the performance of learning algorithms by removing noise and outliers, (IV) a higher efficiency when working on large-scale datasets, and (V) minimizing the labelling cost.

Sample selection methods can be roughly classified into two categories [70]: instance selection and active learning. Instance selection aims at condensing a dataset by filtering noisy and redundant data [2]. Instance selection methods can be categorized into two groups [27]: wrapper methods where the selection criterion is based on the accuracy of a learner, and filter methods where the selection criterion is not based on the results of a learner.

On the other hand, active learning methods process incomplete data, referring to data comprising missing labels, by selecting instances from unlabelled datasets, thus reducing the labelling effort and the cost of training an accurate model [18,52,59]. Nowadays, we find many modern problems in which a huge amount of unlabelled data is available. Sometimes, the labelling process may be subject to little or no cost. However, for many supervised learning tasks, data labelling is a time-consuming process that requires expert handling [59]. Successful applications of active learning include text categorization [30,42,64], image classification [9,31,68], protein structure prediction [6], natural language processing [47,67], information retrieval [69,77], and information extraction [37,60].

Active learning methods can be categorized according to the type of selection strategy (a.k.a. query strategy) used to iteratively select the unlabelled instances [59]. Many different types of query strategies have been proposed in the literature, including: uncertainty-based query [10,38,57,60], version space-based query [1,43,62], expected model change-based query [60,61], expected error reduction-based query [3,44,56], variance reduction-based

* Corresponding author.
E-mail addresses: ogreyes@uco.es (O. Reyes), ahaltalhi@kau.edu.sa (A.H. Altalhi), sventura@uco.es (S. Ventura).

https://doi.org/10.1016/j.knosys.2018.01.033
0950-7051/

query [7,8,58], density-weighted methods [19,46,60], and recently, the Gaussian process-based query has gained increasing attention [78,79].

Active learning is similar to wrapper-based instance selection since they always involve a learning algorithm involved in the process. However, in the active learning process, an annotator (e.g. a human expert) is also required for labelling the selected unlabelled instances. Active learning is an iterative process, that in each iteration a selection strategy selects a set of unlabelled instances, the instances selected are labelled by the annotator, the instances are added to the training set, and the learning algorithm is trained with the new training set.

In this paper, we focus on the active learning paradigm and we aim to study the following problem: *Given n selection strategies that are assessed on m datasets, determine whether the selection strategies differ significantly in performance*. In other words, we aim to study how to statistically compare active learning strategies over multiple datasets.

According to the No Free Lunch theorem, it is not possible to find one algorithm which performs best for all possible problems [72,73]. Consequently, the evaluation of experimental results is considered an essential part of any research, and over the last few years, statistical tests have been increasingly used by authors to validate results and draw conclusions when comparing algorithms. Since the publication of Demšar's work [11], non-parametric statistical tests have been widely used to validate empirical results obtained by algorithms in areas such as machine learning [20,22], data mining [21] and computational intelligence [12,13,21,23]. Non-parametric tests are preferred over parametric tests, due to the absence of strong limitations (normality, independence and homoscedasticity) regarding the kind of data to be analysed [11].

Despite the call made by the machine learning scientific community for a correct statistical analysis of published results, there has not been a rigorous use of statistical tests to compare the performance of the active learning methods. To date, selection strategies have been commonly evaluated by visually comparing learning curves [59]. The visual comparison of learning curves provides a qualitative way to determine whether an active selection strategy outperforms another one. However, the visual comparison of several learning curves can often be very confusing, as the curves may overlap. The visual comparison of learning curves is further complicated when several selection strategies with similar performances are compared over a large number of datasets.

In this work, two comparison approaches are proposed, based on the use of non-parametric statistical tests, to compare active learning methods. The first approach is based on the analysis of the area under learning curve and the rate of performance change. The second approach considers the intermediate results derived from the active learning iterations. To the best of our knowledge, this work is the first attempt at examining how to do statistical comparisons of active learning strategies over multiple datasets. A thorough experimental study was conducted to illustrate the application of the two approaches proposed in this work, evaluating four selection strategies on 26 datasets, showing the usefulness of our proposal for a better comparison of the active learning methods.

The remainder of this paper is arranged as follows: Section 2 shows some basic definitions and the state-of-the-art in the evaluation of the active learning performance. Section 3 presents the two approaches proposed to statistically compare active learning strategies. Section 4 describes the experimental study carried out in this work. Section 5 introduces some guidelines. Finally, Section 6 provides some concluding remarks.

2. Preliminaries

This section describes the basic definitions used throughout this work. Moreover, a review of the state-of-the-art techniques for the evaluation of the active learning performance is carried out.

2.1. Basic definitions

Let us say Φ is the base classifier used in an active learning process, L and U represent the labelled set and unlabelled set, respectively, and θ is a selection strategy that selects a set of unlabelled instances from U using some selection criterion. Active learning is an iterative process, that commonly performs in each iteration the following steps:

1. θ selects a subset of unlabelled instances from U.
2. The selected instances are labelled by an annotator (e.g. a human expert).
3. The selected instances are added to L and removed from U.
4. Φ is trained with the labelled set L.
5. The performance of Φ is tested (optional).

These steps are repeated iteratively. In the active learning literature, several stopping conditions have been proposed. Commonly, the active learning process is repeated k times (number of iterations). However, the performance of the base classifier can be used as stopping criterion when it has attained a certain level. The way of testing the performance of the base classifier depends on the problem studied. The performance of the base classifier is evaluated by using a test set and analysing an evaluation measure.

Let us say L_i and U_i are the labelled set and unlabelled set, respectively, in the ith iteration of the active learning process, and Φ_i is the classifier constructed in the ith iteration using L_i as training set.

Definition 2.1. *Superiority of a classifier*. A classifier Φ_1 is considered superior to a classifier Φ_2, denoted as $\Phi_1 \succ \Phi_2$, if Φ_1 has a better performance than the classifier Φ_2, where both classifiers were assessed under the same conditions.

Definition 2.2. *Ideal selection strategy*. A selection strategy is considered ideal if it is able to select in every iteration a set of unlabelled instances implying the construction of a classifier that it is superior to all classifiers generated in previous iterations.

The following axiom is derived directly:

Axiom 2.1. An active learning process with an ideal selection strategy generates a sequence of classifiers $\Phi_1, \Phi_2 \ldots, \Phi_k$ satisfying $\Phi_k \succ \Phi_{k-1} \succ \ldots \succ \Phi_1$.

An active learning process can be represented as a learning curve which plots the performance attained by the base classifier in every iteration. Fig. 1 represents an active learning process using an ideal selection strategy. Fig. 1a shows a case where the performance of the base classifier was assessed in each iteration by analysing a maximal evaluation measure, whereas Fig. 1b shows a case for a minimal evaluation measure. When a maximal measure is analysed, the higher the value, the better the performance, whereas for a minimal measure the opposite occurs. Without loss of generality, we analysed in this work the case where the base classifier is assessed by using a maximal evaluation measure.

It is always desired that a selection strategy behaves as an ideal selection strategy; i.e. a better classifier would be generated in each active learning iteration. However, this condition is difficult to attain, since the performance of a selection strategy can be biased in several ways by the base classifier used and the characteristics of the unlabelled data. In fact, in some iterations, a selection strategy can select a set of unlabelled instances that possibly could

Please cite this article as: O. Reyes et al., Statistical comparisons of active learning strategies over multiple datasets, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.01.033