Procedia Computer Science

Procedia Computer Science 101, 2016, Pages 45 – 52

YSC 2016. 5th International Young Scientist Conference on Computational Science

New Approaches to Parallelization in Filters Aggregation Based Feature Selection Algorithms

Ivan Smetannikov[1*] and Ilya Isaev[1] and Andrey Filchenkov[1]

[1] *ITMO University, St.Petersburg, Russia.*

ismetannikov@corp.ifmo.ru, isaev@rain.ifmo.ru, afilchenkov@corp.ifmo.ru

Abstract

One of the classical problems in machine learning and data mining is feature selection. A feature selection algorithm is expected to be quick, and at the same time it should show high performance. MeLiF algorithm effectively solves this problem using ensembles of ranking filters. This article describes two different ways to improve MeLiF algorithm performance with parallelization. Our experiments shown that proposed schemes improve algorithm performance significantly and increase feature selection quality.

Keywords: Machine learning, feature selection, rank aggregation, multi-armed bandit, parallel computation, MeLiF, MeLiF+, PQMeLiF, MAMeLiF.

1 Introduction

Almost all business or scientific problems nowadays involve processing huge amounts of data with machine learning algorithms. Due to its universal applicability, machine learning became one of the most promising and researched scientific domains. In particular, it has application in bioinformatics (Bolón-Canedo V. S.-M. N.-B., 2014) (Saeys Y., 2007), as giant amounts of data about gene expression of different organisms are obtained in this field. In order to filter data noise and reduce model complexity, it is necessary to select the most relevant features. Techniques and methods achieving this goal belong are called feature selection. Gene expression data can enable researchers to spot which DNA pieces are responsible for reactions to particular environment change or some internal processes of an organism. The main problem met in processing such data is the high dimensionality of instances. Gene expression datasets often have a high number of features and relatively low number of objects. For a datasets with these properties, it is very hard to build a model that fits data well.

[*] Corresponding author

Peer-review under responsibility of organizing committee of the scientific committee of the 5th International Young Scientist Conference on Computational Science

doi:10.1016/j.procs.2016.11.007

A feature selection algorithm meets several requirements. It is expected to work fast and show good performance. However, no universal algorithm for feature selection exists. *Wrappers* (Kohavi R., 1997) are the family of methods based on searching for an optimal feature subset that maximizes preselected classifier effectiveness. Such a problem statement leads to high performance of found solution. However, the size of search space grows exponentially of the number of features. This fact makes wrapper rarely applicable in bioinformatics, as number of features in datasets could be up to hundreds of thousands. In these cases, another feature selection algorithms known as *filters* (Sánchez-Maroño N., 2007) are used. Filters are based on estimation of feature importance. Filters usually perform worse than wrappers, but they are much faster. A special group of feature selection methods are and *embedded selectors* (Lal, 2006) which uses particular properties of a selected classifier.

Ensembling, which is process of building a combination of several simple algorithms, is a widely used technique in machine learning (Bolón-Canedo V. S.-M. N.-B., 2012). MeLiF algorithm proposed in (Smetannikov I., 2016), applies ensembling to feature selection. This algorithm tries to find such a linear combination of basic ranking filters, which selects the most relevant features of the dataset. Ranking filter itself consists of two separate parts: feature importance measure and cutting rule. Basically, MeLiF tries to tune coefficients of measures linear combination. This process involves classifier training, evaluating and comparing with ranking filters themselves, thus making it is comparatively slow. That is why parallelization can become really handy and helps to improve algorithms computational time.

The simplest parallelization scheme called MeLiF+ is described in (Isaev I., 2016). The main disadvantage of this naive parallelization scheme is that it does not scale well. MeLiF+ calculates feature selection effectiveness using separate thread for each starting point in a search space. When optimization for given point finishes, this thread just stops and its resources stay unreleased therefore they cannot be used for further work. Thus, it is not useful to allocate a lot of resources for this process as most of them will stay unused.

To overcome this problem, it is necessary to use cores of processing server more effectively. While processing, MeLiF visits a lot of points in linear space, so we can process points using a task executor. This research proposes two different approaches to using parallel coordinate descent in building ensembles of ranking filters called PQMeLiF and MAMeLiF. The first algorithm PQMeLiF stores its points that should be processed in a priority queue. The second algorithm MAMeLiF solves parallelization problem by reducing it to multi-armed bandit problem.

The remainder of the paper is organized as follows: Section 2 describes MeLiF algorithm as is, Section 3 contains the proposed parallelization schemes, Section 4 outlines experimental setup, Section 5 contains experiment results, and finally Section 6 contains conclusion.

2 MeLiF

The algorithm tries to find a linear combination of ranking filters that maximizes classification quality using a particular classifier. This detail allows classify MeLiF algorithm as a hybrid of filters and wrappers, which uses filters speed and wrappers focus on resulting performance. The algorithm constructs a linear aggregation of feature importance measures and tries to optimize coefficients in this aggregation thus producing a new feature importance measure. As optimization function it uses F_1 score of chosen classifier after applying new feature selection algorithm for given dataset. This algorithm consists of new feature importance measure, and some cutting rule. Basically, during its work, each aggregation, which is a set of coefficients, is represented wit a point from *N* dimensional space, where *N* is the number of aggregated feature ranking algorithms. MeLiF selects feature subset using coordinate descent not on the feature space, as wrappers usually do, but on the aggregation coefficients space. The number of coefficient is simply the number of filters used in ensemble and it is