Engineering Applications of Artificial Intelligence 74 (2018) 23–34

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

New diversity measure for data stream classification ensembles

Konrad Jackowski

Wroclaw University of Science and Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

ARTICLE INFO

Keywords:
Ensemble classifier
Diversity measure
Data stream classification
Concept drift

ABSTRACT

The diversity of a voting committee is one of the key characteristics of ensemble systems. It determines the benefits that can be obtained through classifier fusion. There are many measures of diversity that can be used in classical decision-making systems which operate in stationary environments. A plethora of algorithms have also been proposed to ensure ensemble diversity. Bagging and boosting are a few of the most popular examples. Unfortunately, these measures and algorithms cannot be applied in systems that process streaming data. Not only must a different implementation be designed for processing fast moving samples in a stream, but the notion of diversity must also be redefined. In this paper it is proposed to assess diversity based on analysis of classifier reactions to changes in data streams. Therefore, two novel *error trend diversity measures* are introduced that compare the error trends of classifiers while processing subsequent samples. A practical application of these measures is also proposed in the form of a novel *error trend diversity driven ensemble* algorithm, where our measures are incorporated into the training procedure. The performance of the proposed algorithm is evaluated through a series of experiments and compared to several competing methods. The results demonstrate that our measures accurately evaluate diversity and that their application facilitates the creation of small and effective ensemble classifier systems.

1. Introduction

Ensemble systems are widely used among researchers who must perform classification tasks. Such systems make a collective decision by combining the responses of several classifiers that form a committee. This approach negates the limitations of single classifiers and elevates classification accuracy for very difficult problems. Naturally, classifier fusion does not guarantee that the ensemble outperforms a single classifier. However, it is widely known that such an approach can be effective when the classifier committee is diversified and its members complement each other (Jain et al., 2000; Zenobi and Cunningham, 2001). In other words, no advantage can be gained by combining several similar predictors. Therefore, two essential points must be addressed: how to measure diversity and how to enforce diversity.

There are several well-known diversity measures, including *coincident failure diversity*, the *disagreement measure*, and *entropy measure* (Tang et al., 2006). Focusing on second essential point, one can refer to the methods for diversity enforcement listed in Brown's survey (Brown et al., 2005). We can divide these methods into four groups: randomising stochastic training algorithms, manipulation of training data, manipulation of architectures, and heuristic methods. All of these measures and methods were originally developed for ensembles working in stationary environments. Therefore, they cannot be easily applied to systems that process streaming data. There are two main reasons for this limitation.

First, streams of data significantly differ from classical learning sets. The last ones consist of fixed number of previously collected samples. Therefore, regardless the number of samples, batch processing of the set can be applied for training. Having got all samples available, they can be repeatedly used until expected convergence of the process is obtained. What more, the time and order of samples acquisition does not matter, therefore they can be shuffled if it help to improve training. On the other hand, data streams by definition have not got fixed size. Sources of streams generate data continuously and samples arrived sequentially. Not only simple batch processing cannot be applied but also samples order must be preserved. And finally, software implementations of data stream classification algorithms must consider the fact that data must be processed in high volumes at fast paces. This consideration can be addressed through the application of online or chunk-based processing modes (Krawczyk et al., 2017). In online model, each sample extracted from the stream is processed separately upon arrival, which allows one to limit memory usage and processing time. Chunk-based algorithms utilise a memory buffer called a chunk to collect extracted samples. When the buffer is filled with a given number of samples, the training procedure begins. The main advantage of this approach is that batch algorithms can be relatively easily adapted.

Second, the notion of diversity must be redefined to reflect the temporal and floating characteristics of data in a stream. Because

E-mail address: konrad.jackowski@pwr.edu.pl.

https://doi.org/10.1016/j.engappai.2018.05.006
Received 10 April 2018; Received in revised form 18 May 2018; Accepted 22 May 2018

this aspect is of key importance, it shall be described in detail. Data streams can be non-stationary, meaning data characteristics change over time. This phenomenon is referred to as *concept drift* (Zliobaite, 2010). It is considered in systems for spam filtering (Delany et al., 2006; Ruano-Ords et al., 2018), financial fraud detection (Hilas and S., 2009), identification of customer preferences (Black and Hickey, 2002), automated production monitoring (Yao and Ge, 2017; Soares and Arajo, 2016), power plant performance modelling (Xu et al., 2017).

The change can affect: (a) a prior probability of classes, when a proportion of each class in the population changes; (b) a class conditional probability, when a conditional distribution of object attributes drifts in feature space; and finally, (c) a posterior probability of classes, i.e. a probability that an object with given attributes belongs to given class. Change of the prior probability is called a virtual concept drift, i.e. a drift which does not affect decision boundaries (Tsymbal, 2004). On the other hand, points (b) and (c) refer to situations when the drift moves the boundaries and, therefore, it is called a real concept drift. Most recent researches on a classification of evolving streams are focused on the real drift. Another important factor, which must be considered, is a dynamic of the drift. A well known taxonomy consists of the following types of the drift: sudden, gradual, and incremental concept drift. The sudden drift takes place when essential changes appear suddenly at a particular moment in time. In the other two cases, changes are spread over time. Regardless the source of changes and their dynamic, appropriate action must be taken to preserve classification accuracy. Usually, replacing classifier with a new one works well in case of the sudden drift. On the other hand, adaptation of the classifier might be a better alternative in case of an evolving changes. The last important question is when to update the classifier. Two main types of approaches can be found in the literature: active and passive approaches (Ditzler et al., 2015). In the first type, a special controller called a drift detector is used to detect concept drift and trigger an appropriate reaction. This may include updating the classifier or rebuilding it from scratch. Passive algorithms continuously update their model when new samples are extracted from a stream (Gomes et al., 2017). This approach is most widely used in ensemble systems, where adaptation typically entails creating a new ensemble member.

Two objectives were defined in the research presented in this paper. First, to design a diversity measure for evolving datastreams, and second, to incorporate new diversity measure in ensemble training procedure. To fulfil the goals two original rules were established

1. Diversity shall be considered to be the ability of ensemble members to react to passing samples in a diversified manner. It is suggested to track the classification errors of elementary classifiers while processing subsequent data samples and comparing their trends. If they display similar trends, it means that their response to the concept drift is similar, meaning they have small diversity. Conversely, error trends moving in opposite directions indicates high diversity.
2. The ensemble shall update its committee continuously by creating new classifiers. However, the composition of the committee shall be controlled by a hybrid target function that aims to minimise ensemble classification error and maximise ensemble diversity.

As the result of implementation of the first rule, two diversity measures were defined. The first one, called the *pair error trend diversity measure*, was designed for pairwise classifier analysis. The *pool error trend diversity measure* evaluates a set or committee of classifiers. Both measures use two subsequent chunks extracted from the stream to evaluate the trend of classifier error. Next, application of the second rule resulted in designing new ensemble classifier called *error and trend diversity driven ensemble* (ETDDE). It is trained using a hybrid target function that operates based on the second rule.

The remainder of this paper is organised as follows. A selection of related works is presented in Section 2. A formal presentation of the diversity measures is provided in Section 3. The details of our novel ensemble training procedure are given in Section 4. Experimental evaluations are presented and discussed in Section 5.

2. Related works

When talking about classifier systems that process streams with concept drift, one should start with the streaming ensemble algorithm (SEA), which is one of the earliest and most well-known algorithms proposed by Street and Kim (2001). In the SEA, the classifier created for the most recent chunk replaces the ensemble member with worse performance in terms of classification accuracy. SEA uses simple majority voting for decision making, but a more sophisticated approach can be found in the accuracy weighted ensemble (AWE), where the weights assigned to classifiers reflect their quality (Wang et al., 2003), or in the accuracy updated ensemble (AUE), where classifier weighting is performed by using a non-linear error function (Brzeziński and Stefanowski, 2011). The well-known bagging and boosting methods inspired several successors, including fast and light boosting (Aboost) (Chu and Zaniolo, 2004), Learn++.NSE (Elwell and Polikar, 2011), and ADWIN bagging (OzaBagAdwin) (Bifet et al., 2009). A change detector is used in the last method, which places it among the active approaches. When a change is detected, the worst classifier in the ensemble is replaced with a new one. Dynamic adaptation to concept changes (DACC) and its improved successor anticipative and dynamic adaptation to concept changes (ADACC) (Jaber et al., 2013) were specifically designed for handling recurring concepts, which are a special case of concept drift. This phenomenon occurs when a past concept appears again. In this particular case, all methods that forget old concepts or discard outdated ensemble members must restart training from scratch when the context emerges again.

All aforementioned methods are well known and often used in comparative analyses. It should be emphasised, however, that an issue of evolving stream classification constantly attracts attention of researchers, as evidenced by a multiplicity of publications. Therefore, to provide a more complete picture of current state of research, we should complete our overview with a presentation of selected latest proposals.

We start with interesting discussion on an issue of choosing appropriate ensemble size (a number of elementary classifiers which form the ensemble). It can be found in Pietruczuk et al. (2017). Authors state that the ensemble shall be updated with new component when this action increases accuracy evaluated not only for recent observations but for the whole data stream. In Wang et al. (2017) a new model-combining methods is presented. It uses constrained and penalised regression, which is especially dedicated for stationary and non-stationary stream processing. It selects data batches relevant to current one and adaptively adjusts a model to the drift. A class imbalance is important phenomenon, which should be also considered when processing streaming data. Usually, researchers solve it by applying variety of resampling strategies as presented in Wang et al. (2015) and Nguyen et al. (2017). An interesting method of classifier adaptation for changes, which does not require labelling, is proposed in Kumagai and Iwata (2017). The adaptation is based on estimating conditional distribution of new features. This approach allows to track the drift and save a time required for labelling. Class-based ensemble for class evolution is a name of an algorithm especially designed to deal with emergence and disappearance of the classes in the streams (Sun et al., 2016). The same problem is also addressed in Mohamad et al. (2018), where proposed method uses an active learning strategy. Adaptation of the active learning paradigm for stream processing, what means a selective sampling and labelling, allows to optimise processing time. For example, in Bouguelia et al. (2016) adaptive uncertainty model is presented. Alternative query by committee active learning strategy presented in Krawczyk and Wozniak (2017) engages a committee in decision making whether to label the sample or not. One of the newest drift detecting method for text streams is proposed in Zhang et al. (2017). Extreme learning machine presented