Signal Processing 148 (2018) 68–77

Contents lists available at ScienceDirect

Signal Processing

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

A hierarchical classification method using belief functions

Daniel Alshamaa[a,*], Farah Mourad Chehade[a], Paul Honeine[b]

[a] Institut Charles Delaunay, ROSAS, Université de Technologie de Troyes, UMR 10300, CNRS, Troyes, France
[b] LITIS Lab, Normandie Université, Université de Rouen, Rouen, France

ARTICLE INFO

Article history:
Received 30 May 2017
Revised 1 February 2018
Accepted 13 February 2018
Available online 15 February 2018

Keywords:
Belief functions
Decision making
Error rate
Hierarchical clustering
Multi-class classification

ABSTRACT

Classification is one of the most important tasks carried out by intelligent systems. Recent works have proposed deep learning to solve the classification problem. While such techniques achieve a very good performance and reduce the complexity of feature engineering, they require a large amount of data and are extremely computationally expensive to train. This paper presents a new supervised confidence-based classification method for multi-class problems. The method is a hierarchical technique using the belief function theory and feature selection. The method predicts, for a new sample input, a confidence-level for each class. For this purpose, a hierarchical clustering approach is adopted to create a two-level classification problem. A feature selection technique is then carried out at each level to reduce the complexity of the algorithm and enhance the classification performance. The belief function theory is then used to combine all information and to give out decisions, by computing the confidence of the sample being in each class. The proposed method has been tested for indoor localization in a wireless sensors network and for facial image recognition using well-known databases. The obtained results prove the effectiveness of the proposed method and its competence as compared to state-of-the-art methods.

1. Introduction

The classification problem is widely tackled in data mining applications. It is stated as follows: given a set of labeled training observations relative to certain features, determine the class label of a new unlabeled data instance. Usually, classification methods include two phases. An offline phase where a model is constructed from training data, and an online phase where a new instance is labeled using the constructed model. The output of such methods is either a discrete label for the new instance, or a numerical score for each class label determining the relative tendency of an instance to belong to different classes. This issue is important in text categorization [1], multimedia applications [2], computer vision [3], medical imaging [4], mobile sensor networks [5], etc.

There are two main types of classification: a flat classification that refers to the standard binary or multi-class methods [6], or hierarchical classification where the classes are classified at each level of a defined dendrogram. Figs. 1 and 2 are examples of such classifications, where $\{a, b, c, d, e, f, g, h\}$ is a set of classes, $C1$, $C2$, and $C3$ are parent nodes, and R is a root node. The parent nodes might be either predefined as a taxonomy or created via hierarchical clustering. In hierarchical models, it is distinguished between local and global classifier approaches. In local classifiers, the hierarchy is taken into account by using a local information perspective. This can be represented by three standard ways: Local classifier per node that trains one binary classifier for each node; local classifier per parent node where, for each parent node, a multi-class classifier is trained to distinguish between its child nodes; and local classifier per level that consists of training one multi-class classifier for each level of the class hierarchy. Although the problem can be tackled using any of the previously described approaches, having a single complex model for all classes reduces the size of the global classification model. This is known as the global classifier approach where one single classification model is built taking into account the whole class hierarchy. The dendrogram is either predefined, or created by means of hierarchical clustering techniques according to similarity metrics.

Although no theoretical evidence or proof whether hierarchical or flat classification models are better [7], experiments throughout previous studies have shown that a better accuracy could be obtained by the former especially for a large number of classes [8,9]. However, a large number of levels in the dendrogram causes slowness in the classification procedure, in addition to the risk of propagating any error in a top level all along the hierarchy [7,9]. In both, flat and hierarchical approaches, classical classification methods such as naive Bayes, neural networks, support vector machines [10–12] can be applied either on the original classes or at each

* Corresponding author.
E-mail addresses: daniel.alshamaa@utt.fr (D. Alshamaa), Farah.chehade@utt.fr (F.M. Chehade), Paul.honeine@univ-rouen.fr (P. Honeine).

https://doi.org/10.1016/j.sigpro.2018.02.021
0165-1684/

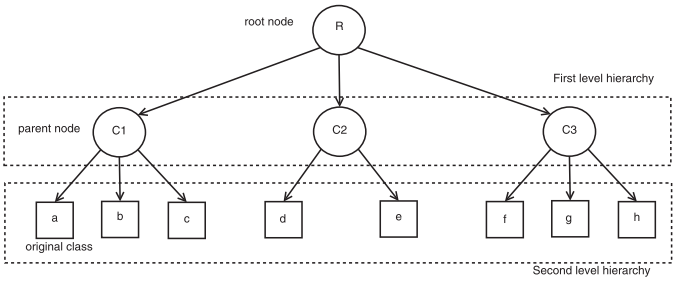

Fig. 1. Hierarchical classification.

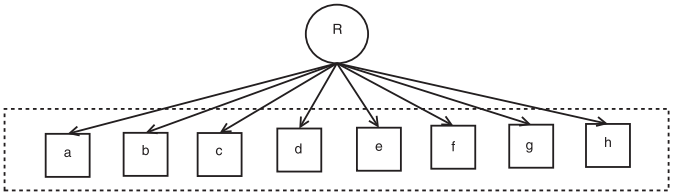

Fig. 2. Flat classification.

level of the hierarchy. Another concept related to the hierarchical approach is the deep learning that uses a cascade of layers for feature extraction and transformation, each layer taking as inputs the outputs of the previous one [13]. Though it has best-in-class performance and reduces the complexity of feature engineering as compared to other solutions in the domain, deep learning requires a large amount of data and is extremely computationally expensive to train. In difference with deep learning, this paper does not tackle feature extraction, rather the work starts once the features are derived. As a consequence, the proposed method can benefit from any feature extraction technique as a preliminary phase, including deep learning.

This paper proposes a confidence-based classification method for multi-class problems. The proposed method is a hierarchical technique, using belief functions and feature selection, described in the following. Given a database of labeled training observations, the classes are merged into clusters using an agglomerative hierarchical clustering method. An optimal level of clustering is selected from the obtained dendrogram, by optimizing the inter- and intra-clusters scatters. The hierarchy is reformed into two levels: the first consisting of the optimal selected clusters, and the second of the original classes in each cluster. Reducing the hierarchy to only two levels decreases considerably the complexity of the method, compared to classic hierarchical methods, with more robustness against error propagation. It also reduces the considered labels at a level, which makes it more efficient than flat techniques. Afterwards, the objective of classification becomes to determine the correct cluster and the correct class at the first and second levels respectively. At each stage, a feature selection technique is applied to choose the best features capable of discriminating between classes and clusters. This creates a framework for the belief function theory (BFT) that associates masses and combines evidence to determine a level of confidence of having the new instance belonging to each class.

The contribution of our work can be summarized as follows. First, the transformation of the problem from a classical flat classification to a two-level hierarchical classification. Second, the feature selection technique. The proposed approach maximizes the discriminative capacity of the ensemble of features at each level of the hierarchy and is consistent with the statistical distributions used to model the observations of the classes. Third, the belief functions framework where masses are assigned to supersets of clusters and classes taking advantage of all available evidence at each level. All assigned masses are then combined to attribute a level of confidence to each original class.

The remainder of the paper is organized as follows. Section 2 is a state-of-the-art that states the problem and defines the concepts needed in the rest of the paper. Section 3 describes the proposed classification approach. Section 4 shows the results of applying the proposed method for facial image recognition and for localization in a wireless sensors network, compared to other well-known state-of-the-art classification techniques. Finally, Section 5 concludes this paper.

2. State-of-the-art

In this section, the classification problem is firstly stated and formulated. Afterwards, the concepts needed in the proposed approach to solve the problem are then introduced.

The proposed classification problem can be formulated as follows. Let

- $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a training dataset of n observations, with $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, p being the number of observations features;
- m be the number of classes and y_i^{cl} denote the class i;
- $L = \{\ell_1, \ldots, \ell_n\}$ be the labels set associated to the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and whose values are taken within $\{y_1^{cl}, \ldots, y_m^{cl}\}$.

The aim of the algorithm is to find a function $\mathbf{h} : \mathbb{R}^p \rightarrow [0, 1]^m$ such that $\mathbf{h}(\mathbf{x}) = (\mathcal{C}f(y_1^{cl}), \ldots, \mathcal{C}f(y_m^{cl}))$, where $\mathcal{C}f(y_i^{cl})$ is the level of confidence of the statement: "$\mathbf{x}$ *belongs to class* y_i^{cl}".

The first step of the proposed approach is merging the classes, once the distributions are defined, using clustering. Clustering aims to organize a set of data into groups called clusters, according to some criteria [14]. Hierarchical clustering builds a hierarchy of clusters or dendrogram driving two strategies: agglomerative or divisive approaches. In the agglomerative or the *bottom up* approach, each observation starts as an independent cluster, and pairs of clusters are merged upon moving up in the hierarchy; whereas in the divisive or *top down* approach, all observations start as one single cluster, and are split upon moving down in the hierarchy [15].

At the end of the developed clustering phase, a two-level hierarchy is obtained. At each level, a feature selection technique is applied to choose the best features. The observations have p components, each one being related to a certain feature, of the set $F = \{f_1, f_2, \ldots, f_p\}$. Feature selection aims at searching for the best subset of the competing $2^p - 1$ candidate subsets of F according to some evaluation function. This can be solved using filter or wrapper method [16]. The filter approach selects feature subsets based on the general characteristics of the data without considering the learning algorithm. Alternatively, the wrapper approach searches for the best subset of features according to an evaluation criterion based on the same learning algorithm. Although the wrapper approach performs better than the filter approach in general, however it is more computationally complex which makes it impractical in many cases [17].

The belief theory, which is also called the DempsterShafer theory or the evidence theory, is a variant of the probability theory where elements are not single points but rather sets or intervals [18]. It is a branch of mathematics that provides an original framework for data fusion based on evidence [19]. In general, the belief function based decision fusion framework mainly includes two phases, mass construction and basic belief assignment (BBA) combination. In the proposed classification method, this is the last step where masses are associated and translated into confidence levels. By taking into account information uncertainty, the proposed method yields several possibilities of classes with different levels of confidence of covering the new observation.