Expert Systems with Applications 42 (2015) 5591–5606

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Multithreshold Entropy Linear Classifier: Theory and applications

Wojciech Marian Czarnecki *, Jacek Tabor

Faculty of Mathematics and Computer Science, Jagiellonian University, prof. Stanislawa Lojasiewicza 6, 30-348 Krakow, Poland

ARTICLE INFO

Article history:
Available online 18 March 2015

Keywords:
Classification
Renyi's entropy
Density estimation
Multithreshold classifier

ABSTRACT

This paper proposes a new multithreshold linear classifier (MELC) based on the Renyi's quadratic entropy and Cauchy–Schwarz divergence, combined with the adaptive kernel density estimation in the one dimensional projections space. Due to its nature MELC is especially well adapted to deal with unbalanced data. As the consequence of both used model and the applied density regularization technique, it shows strong regularization properties and therefore is almost unable to overfit. Moreover, contrary to SVM, in its basic form it has no free parameters, however, at the cost of being a non-convex optimization problem which results in the existence of local optima and the possible need for multiple initializations.

In practice, MELC obtained similar or higher scores than the ones given by SVM on both synthetic and real data from the UCI repository. We also perform experimental evaluation of proposed method as a part of expert system designed for drug discovery problem. It appears that not only MELC achieves better results than SVM but also gives some additional insights into data structure, resulting in more complex decision support system.

1. Introduction

Linear classifiers (SVM, perceptron, LDA, logistic regression) are one of the most commonly used models in modern expert systems. Despite their simplicity, their applicability span among many disciplines from human action recognition (Chaaraoui, Padilla-López, Climent-Pérez, & Flórez-Revuelta, 2014) to credit risk prediction (Danenas & Garsva, 2015). In this paper we analyze strongly related class of models, namely multithreshold linear classifiers, which change the typical decision rule from

$$\operatorname{sign}(v^T x - b) \tag{1}$$

to its multithreshold equivalent

$$\prod_{j=1}^{k} \operatorname{sign}(v^T x - b_j).$$

Introduced Multithreshold Entropy Linear Classifier (MELC) is able to build such model by simultaneously maximizing resulting multiple margins[1] similarly to the Support Vector Machines and to Two ellipsoid Support Vector Machines (Czarnecki & Tabor, 2014), which maximize margin defined by two different metrics. Furthermore it is a model internally designed to deal with unbalanced datasets which normally require careful adaptation of the existing methods (Thammasiri, Delen, Meesad, & Kasap, 2014).

By plugging such a model into the expert system one can benefit from some important features. First, the learned model is extremely simple, so can be efficiently used in numerous applications, including real-time systems. Second, multithreshold nature of the model (see for example Fig. 1) results in the significant increase in the classification accuracy. It also gives additional insight into data structure through performing the clustering which enables system's user to identify underrepresented subclass of samples. One could use this feature to perform active learning (Zhang, Wen, Wang, & Jiang, 2014) and query labels of such samples. As a real-world test of the concept, we apply MELC as a module for the decision support system responsible for finding new drug candidates using the so called ligand based approach. It is a dual method to the one taken by for example Quevedo, De Paris, Ruiz, and de Souza (2014), where the protein structure is used to perform a valid prediction. Performed experiments show significant increase in the classification accuracy over analogous system using Support Vector Machines as well as Perceptron algorithm.

These advantages are obtained at the cost of dealing with an computationally expensive, non-convex function to optimize. Consequently, there are many local optima of MELC objective and one cannot guarantee global convergence. As a result, systems based on this model require more time and effort to train, but as the result give a robust classification rule, with good accuracy

* Corresponding author. Tel.: +48 12 664 7556.
E-mail addresses: wojciech.czarnecki@uj.edu.pl (W.M. Czarnecki), jacek.tabor@uj.edu.pl (J. Tabor).

[1] More details are given in the "Theory: largest margin classifiers" section.

http://dx.doi.org/10.1016/j.eswa.2015.03.007
0957-4174/

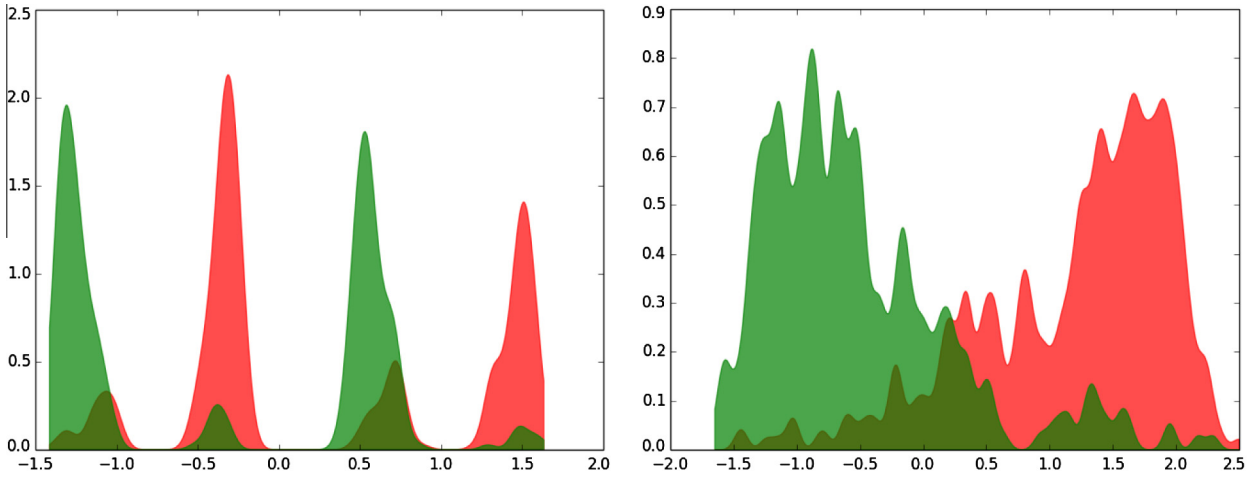

Fig. 1. Comparison of the australian dataset modeled by MELC (on the left) using 3-threshold linear classifier and by SVM (on the right).

and additional advantages absent in both linear models and even kernelized methods (see Evaluation Section for more details).

Let us now briefly describe the contents of the paper. After outlining general idea of the model and short analysis of related work we show the basic properties of the objective function including its scale invariance and solutions for normally distributed data. Next, we prove that proposed model maximizes the margins' sizes of multithreshold linear classifier and what plays the regularization role. Then we proceed to some practical considerations regarding optimization procedure, its implementation and possible drawbacks. We conclude with the evaluation based on both synthetic and real datasets, including proposed expert system for drug discovery.

2. General idea

The linear classification is important as it has the advantage of small VC dimension and as a result are less prone to overfit than more complex models (Vapnik, 2000). The same ideas can be seen behind the neural networks and their modifications like Extreme Learning Machines (Huang, Wang, & Lan, 2011) or Deep Learning (Hinton, Osindero, & Teh, 2006), where the activation of the single neuron is given by (1), while the role played by it in the whole decision process is usually given by

STEP 1: calculate $v^T x$,
STEP 2: make decision based on the sign of $v^T x - b$.

Although the linear classification is usually very efficient, even for the simple sets in $\mathbb{R}$, like $+\,-\,+$, see Fig. 2, we cannot obtain sufficient classification results. This led to the need for kernelization procedure (Cortes & Vapnik, 1995).

Our postulate is that by applying the second step we often lose some of the information given by the first one – observe that both in $+\,-\,+$ or XOR case we can make sufficiently good classification decision based on the knowledge of the value of $v^T x$ (for well chosen v), see Fig. 6. One can therefore ask why we do not use the additional information? One of the possible answers lies in the fact that most classification methods, like SVM, aim at building a "large" linear margin between classes, which in a natural way leads to the single-threshold decision boundary.

Thus there appears a natural question if we can construct a classification method which would find the projection $x \to v^T x$ able to directly deal with more complex classification cases like $+\,-\,+$ and XOR. The problem in fact splits into two – how to find the right $v \in \mathbb{R}^d$ and how to make the proper classification decision in $\mathbb{R}$. The answer for the second question is given by multithreshold linear classifiers (Cao, Cuevas, & Gonzalez Manteiga, 1994), where instead of decision based on the split of $\mathbb{R}$ into $(-\infty, b)$ and $[b, \infty)$ the division into finite number of intervals is allowed[2].

The answer to the first question is nontrivial, and in our opinion there could be many reasonable solutions. In this paper we have decided to base the decision on entropy-based divergence measure (Principe, Xu, & Fisher, 2000c). We have chosen the *Renyi's quadratic entropy* H_2 and *Renyi's quadratic cross entropy* (Principe et al., 2000c) $H_2^\times$

$$H_2^\times(f,g) = -\log \int fg$$
$$H_2(f) = H_2^\times(f,f) = -\log \int f^2,$$

as well as connected *Cauchy–Schwarz divergence*

$$D_{cs}(f,g) = \log \int f^2 + \log \int g^2 - 2\log \int fg = 2H_2^\times(f,g) - H_2(f) - H_2(g). \quad (2)$$

Our reasons behind such a choice are the following:

- Renyi entropy and the Cauchy–Schwarz divergence are easily computable and the exact formulas for the Gaussian mixtures are known (this allows the use of gradient methods in our optimization problem, see Practical Considerations Section),
- the Cauchy–Schwarz divergence is affine transformations invariant in terms of input data transformation,
- D_{cs} has nice theoretical properties, as the maximization of $H_2^\times$ leads to the maximization of the multi-threshold boundary[3], while the part consisting of H_2 adds the regularizing term, see section: "Theory: largest margin classifiers".

From the practical point of view, we first project the data by v^T onto $\mathbb{R}$, and apply there the classical kernel density estimation given for the dataset $P \subset \mathbb{R}$ by

$$[\![P]\!]_\sigma = \frac{1}{|P|}\sum_{p\in P}\mathcal{N}(p,\sigma^2), \quad (3)$$

[2] This type of classification can be obtain in particular by the density based classifiers in $\mathbb{R}$.
[3] To some extent we obtain multi-threshold analogue of large margin classifier.