Cluster-Based One-Class Ensemble for Classification Problems in Information Retrieval

Nedim Lipka
lipka.nedim@gmail.com

Benno Stein
benno.stein@uni-weimar.de

Maik Anderka
maik.anderka@gmail.com

Bauhaus-Universität Weimar
99421 Weimar, Germany

ABSTRACT

A number of relevant information retrieval classification problems are one-class classification problems at heart. I.e., labeled data is only available for one class, the so-called target class, and common discrimination-based classification approaches, be them binary or multiclass, are not applicable. Achieving a high effectiveness when solving one-class problems is difficult anyway and it becomes even more challenging when the target class data is multimodal, which is often the case. To address these concerns we propose a cluster-based one-class ensemble that consists of four steps: (1) applying a clustering algorithm to the target class data, (2) training an individual one-class classifier for each of the identified clusters, (3) aggregating the decisions of the individual classifiers, and (4) selecting the best fitting clustering model. We evaluate our approach with four datasets: an artificially generated dataset, a dataset compiled from a known multiclass text corpus, and two datasets related to one-class problems that received much attention recently, namely authorship verification and quality flaw prediction. Our approach outperforms a one-class SVM on all four datasets.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*clustering, information filtering*

Keywords: One-class Classification, Ensemble, Clustering

1. ONE-CLASS CLASSIFICATION IN IR

In a one-class problem one is given information of the target class only. The task is to define a boundary that encloses as many target objects as possible while minimizing the chance of accepting objects from outside the target class, so-called outliers [8]. An example for a one-class problem is authorship verification [4], where we are given writing examples for a single author T, and we are asked whether a text of unknown authorship was written by T as well. Notice that, despite the fact that a sheer endless number of outliers are at our disposal, we are not able to define a closed outlier class with texts from other authors. Specialized one-class classifiers shall cope with this setting; however, there is no cure-all for one-class problems, and additional constraints may render the classification task even more challenging. Two of which are pretty common in information retrieval: a highly diverse target class, i.e., a target class with a complex, multimodal data distribution, and, the presence of noise in the application situation of the one-class classifier. In this paper we present a cluster-based one-class ensemble that is able to effectively alleviate both problems.

Combining binary or multiclass classifiers within an ensemble has been proven to increase accuracy in many applications, compared to the best individual classifier in the ensemble [3]. With regard to one-class problems only few ensemble strategies have been proposed, which can be distinguished into two categories. First, approaches that divide the feature space and train individual one-class classifiers on the different feature subsets [9, 5]. Second, approaches that divide the target class data and train individual one-class classifiers on the different object subsets [10, 7]. Our approach belongs to the second category and is related to the work of Wang et al. [10], who employ an agglomerative hierarchical clustering strategy in order to partition target class data. By contrast, we employ a more suitable clustering technology, apply our approach to real-world information retrieval tasks, and empirically demonstrate its advantage compared to a common one-class SVM.

2. CLUSTER-BASED ONE-CLASS ENSEMBLE

Let $\mathbf{D}_T = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denote the set of n feature vectors when given n target objects. The construction of the ensemble classifier happens within four steps:

Step 1: Clustering. $\mathbf{D}_T$ is clustered using the k-means algorithm [2]. The algorithm aims to find an exclusive clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, $C_j \subseteq \mathbf{D}_T$, $j \leq k \leq n$, such that the variance in each cluster C_j is minimized. The only free parameter in this step is the cluster number k. Figure 2 shows an example clustering.

Step 2: One-Class Classification. For each cluster $C_j \in \mathcal{C}$ a one-class SVM $c_j : \mathbf{x} \rightarrow \{1, -1\}$ is learned. Similar to a binary SVM a kernel function ϕ is used to map the objects into a higher-dimensional space. The goal is to find a maximum-margin hyperplane in the kernel space that separates $(1 - v) \cdot n$ target objects from the origin. This is formulated as optimization problem [6]:

$$\min_{\mathbf{w},\xi,\rho} \tfrac{1}{2}||\mathbf{w}||^2 + \tfrac{1}{v \cdot n} \sum_{i=1}^{n} \xi_i - \rho$$

$$\text{s.t.} \quad (\mathbf{w} \cdot \phi(\mathbf{x}_i)) \geq \rho - \xi_i \text{ with } \xi_i \geq 0, \ i = 1, .., n,$$

where $\mathbf{w}$ denotes the normal vector of the hyperplane, ρ the margin, and ξ_i the slack variables. The value $v \in (0, 1]$ is specific to a one-class SVM: it defines the fraction of target objects outside the target class and controls the number of

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

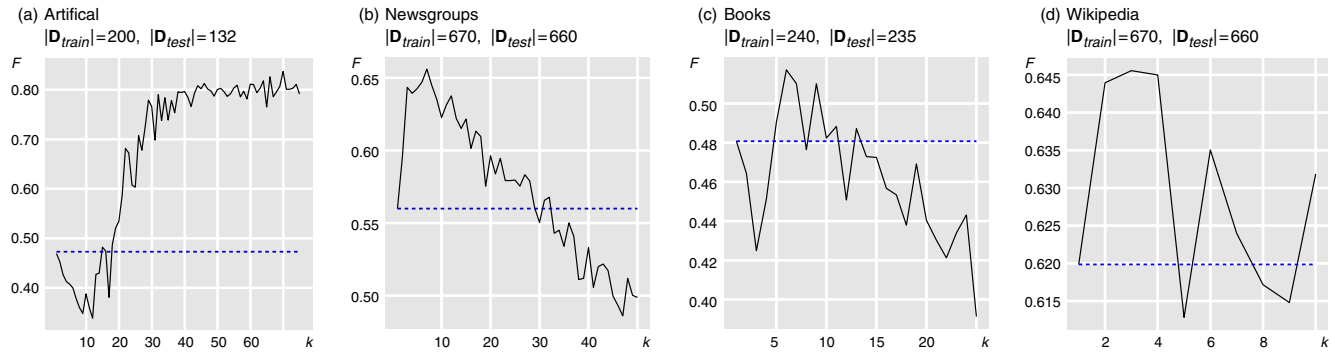

Figure 1: Effectiveness of the cluster-based one-class ensemble approach in terms of F-measure over cluster number k on four datasets. The dotted line shows a one-class SVM, trained with all target class data.

support vectors. The decision function c_j is of the following form:

$$c_j(\mathbf{x}) = sign((\mathbf{w} \cdot \phi(\mathbf{x})) - \rho)$$

Step 3: Aggregation. The ensemble classifier e_k combines the decisions of the k single classifiers for a vector $\mathbf{x}$ as follows:

$$e_k(\mathbf{x}) = \begin{cases} 1 & \text{if } \exists\, c_j(\mathbf{x}) > 0, j = 1, \ldots, k \\ -1 & \text{otherwise.} \end{cases}$$

Step 4: Model Selection. The clustering parameter k runs from 1 to l, and altogether $l(l+1)/2$ one-class SVMs are constructed. We choose the classifier e_k that performs best on holdout validation data in terms of the classification error.

3. ANALYSIS AND RESULTS

We use a random subset $\mathbf{D}_{train} \subset \mathbf{D}_T$ for the training phase; the test set $\mathbf{D}_{test}$ is comprised of a balanced number of outliers and examples from $\mathbf{D}_T \setminus \mathbf{D}_{train}$. Each experiment is repeated 15 times. The effectiveness of the classifier is reported as averaged F-measure and for varying values of the cluster number k. In all experiments a one-class SVM with a non-linear RBF kernel is used. The parameters of classifier c_j are optimized on the respective training set $C_j \subseteq \mathbf{D}_{train}$; clusters with less than five elements are discarded. Figure 1 illustrates the results on four different datasets: artificially created objects with three clusters (plot in Figure 2), documents from the 20 Newsgroups dataset with category "computer" in the role of the target class, books from different authors for which the authorship is to be verified [4], and Wikipedia articles tagged with certain quality flaws that are to be detected [1]. All documents are represented under a vector space model with a tf-idf weighting except for Wikipedia articles where quality-specific features [1] are employed. On all four datasets our approach outperforms a one-class SVM that has been trained with all target objects, or equivalently, where $|\mathcal{C}| = 1$. Table 1 summarizes the achieved improvements (significant) over the baseline. Note that our approach is a meta-classifier. Thus, arbitrary clustering and one-class classification technology can be applied. Both k-means and density estimation can be computed in parallel, so our approach is applicable on huge data as well.

Table 1: Percentage of improvement over the baseline for each dataset and for the optimum cluster number k.

Artificial	Newsgroups	Books	Wikipedia
+34% ($k = 70$)	+10% ($k = 8$)	+3.5% ($k = 6$)	+2.6% ($k = 3$)

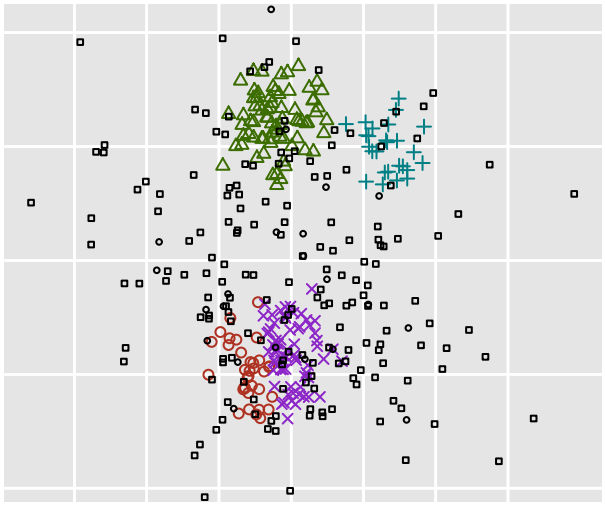

Figure 2: Artificially generated dataset with a diverse target class, constructed as a mixture of Gaussians with one mode for the outlier class and three for the target class. The black squares are outliers, all other objects belong to the target class. The shape and color of the objects indicate a k-means clustering with $k = 4$.

4. REFERENCES

[1] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content: The case of wikipedia. In *Proc. of SIGIR*, 2012.

[2] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc. Series C (Applied Statistics)*, 1979.

[3] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998.

[4] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proc. of ICML*, 2004.

[5] R. Perdisci, G. Gu, and W. Lee. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In *Proc. of ICDM*, 2006.

[6] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comp.*, 2001.

[7] A. Shieh and D. Kamm. Ensembles of one class support vector machines. In *Proc. of MCS*, 2009.

[8] D. M. J. Tax. *One-Class Classification.* PhD thesis, Delft University of Technology, 2001.

[9] D. M. J. Tax and R. P. W. Duin. Combining one-class classifiers. In *Proc. of MCS*, 2001.

[10] D. Wang, D. S. Yeung, and E. C. C. Tsang. Structured one-class classification. *IEEE Trans. Syst. Man. Cybern. Part B Cybern.*, 2006.