Selected Random Subspace Novelty Detection Filter

Fatma Hamdi

CGI BC
Paris la Défense 92097, France
firstname.secondname@cgi.com

Abstract. In this paper we propose a solution to deal with the problem of novelty detection. Given a set of training examples believed to come from the same class, the aim is to learn a model that will be able to distinguich examples in the future that do not belong to the same class. The proposed approach called Selected Random Subspace Novelty Detection Filter ($SRS - NDF$) is based on the bootstrap technique, the ensemble idea and model selection principle. The $SRS - NDF$ method is compared to novelty detection methods on publicly available datasets. The results show that for most datasets, this approach significantly improves performance over current techniques used for novelty detection.

1 Introduction

The task of novelty detection consists of identifing a new data that differs from those used in the training phase of a machine learning system. Several important works in the machine learning literature have addressed the issue of novelty detection and broad reviews of the subjet can be found in [1] and [2]. Novelty detection is an important learning problem, the basic idea is to build a decision rule that distinguishes *normal* from *novel* pattern. Since we can never train a machine learning system on all possible data that the system may deal with, it becomes important that it is able to detect *new* data. In order to overcome the limitations of individual learning algorithms and face the necesstiy of high classification performance specially in some critical domains, many researchers have been interested in ensemble methods. The aim of these techniques is to produce and combine multiple classifiers. Bagging [4], Boosting [5], random forest [8] and their variants are the most popular examples of this methodology. Bagging, a name derived from bootstrap aggregation, was the first effective method of ensemble learning and is one of the simplest methods of arching.

Generally the ensemble methods [16] work on two steps. The first one is the production of homogeneous or heterogeneous models. Models built from the same learning algorithm are called homogeneous and others that derive from running different learning algorithms on the same data set are called heterogeneous. The second step is the agregation of the models. Several techniques here include voting, weighted voting, selection and stacking. The ensemble selection algorithms was proposed to determine the good sub ensemble of classifiers. In supervised

M. Lee et al. (Eds.): ICONIP 2013, Part I, LNCS 8226, pp. 344–351, 2013.

classification, it is known that selective classifier ensembles can always achieve better results compared to traditional ensemble methods [16]. The ensemble selection also called in the litterature ensemble prunning, ensemble overproduce or choose paradigm, consists in choosing a subset of l classifiers from the initial ensemble of size L ($l \leq L$). The selection of classifiers is based on predefined criteria. Generally the proposed approaches rearrange the initial ensemble and select a subset of ensemble members from the sorted list.

In this paper we present a learning model called Selected Random Subspace Novelty Detection Filter ($SRS - NDF$). It is a new approach to novelty detection, wich involves learning from only one class of traning example. We have a sample from one distribution *normal samples*, and our purpose is to differentiate between these normal examples and those that do not appear to come from the same distribution (*novelty*). The $SRS-NDF$ is an extension of our novelty detection model ($RS - NDF$) proposed in [17]. $SRS - NDF$ is based on the bootstrap technique, the ensemble idea [16] and model selection principle [18]. The methodologies for the production and combination of multiple predective models is a very active research area and it is commonly referred to ensemble method. The advantages of these methods are the improvement of the models estimation and the potential improvement of the scalability of their learning algorithms. The main idea of our approach is to perform classifier selection from an initial pool of filters [3] obtained with the ($RS - NDF$) algorithm. The proposed method works by evaluating the qualities of all obtained filters in terms of pertinence. Next we use the scree test [19] to choose the part of pertinent filters to build our final system.

The rest of the paper is organized as follow: Section 2 introduces the basic concepts of the Selected Random Subspace Novelty Detection Filter. Section 3 describes the databases and the experimental protocol. In section 4 we show validation results and their evaluation. Conclusion is given in section 5.

2 Selected Random Subspace Novelty Detection Filters

2.1 Principle of the Kohonen and Oja's Novelty Filter

In 1976, Kohonen and Oja [3] introduced an orthogonalising filter which extracts the parts of an input vector that are, new, with respect to previously learned patterns. This is the desired functionality of a novelty filter. The novelty filter shows the novelties in an input pattern with respect to previously learned patterns. Furthermore, the novelty filter can distinguish the missing parts from the added parts in the input pattern with respect to the previously learned patterns. The construction of the filter is based on Greville's theorem [15]. This theorem gives a recursive expression to estimate the transfer function of the network as follows:

$$\Phi_k = \Phi_{k-1} - \frac{\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T}{\| \tilde{\mathbf{x}}_k \|^2} \tag{1}$$

where $\mathbf{x}_k = [x_1, x_2, ..., x_d]^T$ is a d-dimensional vector from the reference data matrix; $\tilde{\mathbf{x}} = \Phi_{k-1}\mathbf{x}_k$ represents the orthogonal projection of the vector $\mathbf{x}_k$ in the subspace of novelty (Φ_{k-1}). This subspace is orthogonal to the space defined by the first $k-1$ reference data and $\Phi_0 = I$.

An interesting alternative approach was given by Kassab and al. [6], [7], that introduces the identity matrix in the learning formula for considering separately all training examples, and consequently all their features. During the learning phase, features which frequently appear in the training examples become more and more habituated as compared to the less frequent ones. This helps to more discriminate the relevant and irrelevant examples.The new learning rule is then defined as:

$$\Phi_k = \mathbf{I} + \Phi_{k-1} - \frac{\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T}{\| \tilde{\mathbf{x}}_k \|^2} \tag{2}$$

where $\tilde{\mathbf{x}}_k = (\mathbf{I} + \Phi_{k-1})\mathbf{x}_k$ and Φ_0 is zero, or null matrix. The work described in this paper uses this new learning rule.

For the novelty detection problem, two proportions can be computed:

- *Novelty proportion*: this measure quantifies the novelty of an input data with respect to data that has been previously seen during the training.

  $$N_{\mathbf{x}_i} = \frac{\| \tilde{\mathbf{x}}_i \|}{L \times \| \mathbf{x}_i \|} \tag{3}$$

  where L is the number of examples used for the training.
- *The habituation proportion*: this measure calculates the similarity of an example with the previously learned one: $H_{\mathbf{x}_i} = 1 - N_{\mathbf{x}_i}$.
  This proportion could be considered as the classification score of an example x_i. It indicates the probability that x_i belongs to the novel class.

To determine a detection threshold for each filter, the following principle was used:

- Scores (output's filter) attributed to the learning data can be used as a good indicator of the scores of data which can be positive and which are easy to detect because they are strongly similar to the data used for the learning. Consequently, the average of these scores can be admitted as a higher limit for the detection threshold.
- The scores attributed to available data for learning before their use, can be used as a good indicator of the scores of data which are positive but which are less easy to detect. Consequently, the average of these scores can be admitted as a lower limit for the detection threshold.

2.2 SRS-NDF Algorithm

In this section, we present an extention for the novelty detection algorithm $RS-NDF$. The $RS-NDF$ [17] approach uses multiple versions of a training set by using a double bootstrap, i.e. sampling with replacement on examples and sampling without replacement on features. Each of these data sets is used to train a different NDF model. The $RS-NDF$ is then an ensemble of NDF, induced from bootstrap samples of the training data, using random features and examples selection in the model induction process. Prediction is made by aggregating (majority vote) the predictions of the ensemble to create a single output. Our method, called $SRS-NDF$, consists in selecting the ensemble members from a set of individuals filters wich gives better results, in terms of pertinence, than the original ensemble. $SRS-NDF$ belongs to the model selection approaches that reorder the original ensemble members based on pertinence criteria and select a subset of ensemble members from the sorted list using the scree test. $SRS-NDF$ works by evaluating the "index of balanced accuracy" and "diversity" of the filters in the $RS-NDF$ and selecting the promising filters. The final solution is achieved by combining all the selected filters from the original ensemble. To study the pertinence of each filter fl we used the following function:

$$Pertinence_{\alpha}(fl_1) = \alpha \times IBA_{\alpha}(fl_i) + (1-\alpha) \times mean(Div_{\alpha}(fl_i); fl_i); i \in [1, NF] \tag{4}$$

Where $IBA_{\alpha}(fl_i)$ and $Div_{\alpha}(fl_i)$ stands respectively for the index of balanced accuracy [20] and the diversity of the filter fl_i. The index of balanced accuracy is defined by the product of two terms Dominance and Gmean [20]. The first term is a simple measure evaluating the correct predictions of each filter, the second term is the geometric mean of accuracies measured separately on each class. The asset of IBA measure is the ability to distinguish the contribution of each class for overall performance. This means that different combinations of the true positive rate and the true negative rate don't provide the same IBA value and gives the pertinence to the positive class. This measure computes the area of a rectangular region in a two-dimensional space called "Balanced Accuracy Graph". The diversity of two classifiers consist on assigning differents labels to the same examples. Many measures have been proposed to quantify the diversity between two classifiers. In our work, we propose to use the mean Frobenius distance between the transfer matrix of fl_i and the other filters in $RS-NDF$. The coefficient α, $0 \leq \alpha \leq 1$, is a control parameter that balances the accuracy and diversity. The pertinence is then defined as a weighted combination of the diversity and accuracy. Once the pertinence have been calculated for a given α, we then used an established statistical method, scree test, to select the most important filters. This statistical test was initially developed to provide a visual technique to select eigenvalues for principal components analysis.

The basic idea is to generate a curve associated with eigenvalues, allowing random behavior to be identified. The number of components retained is equal to the number of values preceding this "scree". Often the "scree" appears where the slope of the graph changes radically. We therefore needed to identify the point of maximum deceleration in the curve.

Assuming that we have pertinence vector $\mathbf{Per}_k = (Per_{1k}, Per_{2k}..., Per_{jk}, ..., w_{nk})$. Thus we have to process the steps: **Scree Test Acceleration Factor**

1. Sort the pertinence in descending order $\mathbf{Per}_k$. The we obtain a new order $\mathbf{Per}_k = (Per^{1}_{..}, Per^{2}_{..}..., Per^{i}_{..}, ..., Per^{n}_{..})$; where Per^i indicates the index order.
2. Compute the first difference $df_i = Per^{i}_{..} - Per^{i+1}_{..}$;
3. Compute the second difference (acceleration) $acc_i = df_i - df_{i+1}$
4. Find the scree: $\max_i (abs(acc_i) + abs(acc_{i+1}))$
5. Cut and consider all the filters until the scree; (use initial indices of filter before sorting)

The SRS-NDF learning algorithm is shown below:

Algorithme 1. Selected Random Subspace Novelty Detection Filter

```
Repeate for all $\alpha \in [0, 1]$
  1. Construct the $RS - NDF$ with $NF$ filters.
  2. Calculate the IBA of filters $= IBA_\alpha(1), ..., IBA_\alpha(NF)$.
  3. Calculate the diversity of filters $= Div_\alpha(1), ..., Div_\alpha(NF)$.
  4. Calculate the pertinence of filters $= Pertinence_\alpha(1), ..., Pertinence_\alpha(NF)$.
     $Pertinence_\alpha(fl_1) = \alpha \times IBA_\alpha(fl_1) + (1 - \alpha) \times meanDiv_\alpha(fl_1)$;
     $fl_i; i \in [1, NF]$
  5. Select the subset of models using the ScreeTest $= SelectedFilters_\alpha$
  6. Calculate the IBA of the selected filters $= IBA_{SelectedFilters_\alpha}$
  7. $\alpha = \alpha + 0,1$
Until $\alpha = 1$

  – Select the subset with the best value of $IBA$
  – Aggregate the predictions of the selected ensemble and save the novelty
    detection results in $D$.
```

3 Experiments

3.1 Databases Description

We performed several experiments on many relevant data sets: *Spectf*, *Waveform*, *Wine* and *Yeast* from the UCI repository [9] , and Oil [11]. These data sets are summarized in table 1. Since all of the datasets are for binary or multi-class classification problems, they were transformed into novelty detection context. We chose randomly, from each data set, a class as the novelty class and collapsed the rest of the classes into one, and use the modified datasets to evaluate the performance of our approach.

Table 1. Data Set summary

Dataset	Dimension	Size	Size of Novelty class
Oil	48	937	41
Spectf	44	187	15
Waveform	21	5000	1647
Wine	13	178	59
Yeast	8	1484	244

3.2 Performance Measurement and Experimental Protocol

To evaluate the performance of our approach we used several metrics such as true negative rate (Acc-), true positive rate (Acc+)(recall), precision, F-measure and G-mean. These metrics have been widely used for comparaison.

For each dataset, the performance of the classifier ensemble obtained by $SRS-NDF$ was compared to the unpruned filter ensemble obtained from $SR-NDF$, the basic model NDF and the traditional novelty detection methods : The one class Support Vector Machines (SVM) [14], The Principal Components Analysis (PCA) [13], The auto-associative Multi Layers Perceptron (MLP) [12].

We also used the Area under the ROC Curve (AUC) [10]. There are several methods to estimate the area under the ROC curve. In the case of binary classification, the balanced AUC_b is defined as:

$$AUC_b = \frac{Acc - +Acc+}{2} \tag{5}$$

4 Results

For each database, the six approaches have been used and their results have been evaluated in terms of the six performances metrics. The table 2 below shows the performance of the different algorithms on all data sets.

Based on the table above, some conclusions can be drawn.

The results of *wine* data set shows that $SRS - NDF$ is the superior approach to novelty detection. Our approach outperforms the results obatained by all others methods on all metrics. For $Waveform$ dataset, $SRS - NDF$ shows a great improvement over all algorithms on $Acc-$, $Acc+$, AUC_b and $G - mean$ metrics. Except on *precision* and $F - measure$, $RS - NDF$ gives the better results. For $Spectf$ dataset, our algorithm outperforms the other methods on $Acc+$ and $F - measure$ but gives a slightly inferior results on AUC_b and $G - mean$. Genarally our proposed approach $SRS - NDF$ gives better results compared to $RS - NDF$ on $Acc+$ and $F - measure$ metrics. The $Acc+$ measure represents the models capacity to detect the novelty class. We chose this metric in purpose to show the good capacity of $SRS - NDF$ to detect the novelty class. As we can see, $SRS - NDF$ shows excelent results comparing to $RS - NDF$ on all datasets. Considering the $F - measure$, our algorithm gives favorable improvement over $RS - NDF$ on Oil, $Wine$ and $Spectf$ datasets. For $Waveform$ data, $RS - NDF$ outperforms our proposed approach. This metric, that combines

Table 2. Performance comparison on all data sets

Wine	$Acc-$	$Acc+$	$Prec$	$F-$	AUC_b	$G-$
	$(Recall)$			$measure$		$mean$
MLP	0,78	0,68	0,83	0,81	0,73	0,73
ACP	0,60	0,69	0,80	0,68	0,65	0,64
SVM-1C	0,68	0,76	0,81	0,73	0,72	0,71
NDF	0,84	0,78	0,88	0,86	0,81	0,81
RS-NDF	0,87	0,85	0,92	0,89	0,86	0,86
SRS-NDF	**0,90**	**0,93**	**0,93**	**0,91**	**0,92**	**0,92**
Waveform						
MLP	0,67	0,35	0,51	0,58	0,51	0,48
ACP	0,68	0,35	0,68	0,68	0,51	0,49
SVM-1C	0,88	0,22	0,70	0,78	0,55	0,44
NDF	0,68	0,57	0,77	0,72	0,63	0,62
RS-NDF	0,85	0,53	**0,79**	**0,82**	0,69	0,67
SRS-NDF	**0,90**	**0,60**	0,70	0,78	**0,75**	**0,70**
Spectf						
MLP	0,60	0,78	0,42	0,49	0,69	0,68
ACP	0,51	0,82	0,43	0,47	0,67	0,65
1-SVM	0,73	0,74	0,43	0,54	0,74	0,74
NDF	**0,76**	0,75	0,45	0,56	**0,76**	**0,76**
RS-NDF	0,69	0,79	**0,47**	0,56	0,74	0,74
SRS-NDF	0,64	**0,84**	0,44	**0,58**	0,74	0,74
Yeast						
MLP	0,77	0,23	0,84	0,80	0,50	0,39
ACP	0,68	0,66	0,91	0,78	0,67	0,67
SVM-1C	0,90	0,13	0,84	0,87	0,51	0,34
NDF	0,88	0,19	0,85	0,86	0,54	0,41
RS-NDF	**0,94**	0,29	0,87	**0,90**	0,62	0,52
SRS-NDF	0,93	**0,53**	**0,93**	**0,90**	**0,72**	**0,70**
Oil						
MLP	0,96	0,15	0,96	0,96	0,55	0,38
ACP	0,94	0,24	0,96	0,95	0,59	0,48
SVM-1C	0,90	0,35	**0,97**	0,93	0,62	0,56
NDF	0,91	0,37	**0,97**	0,94	0,64	0,58
RS-NDF	0,94	0,22	0,96	0,95	0,58	0,46
SRS-NDF	**0,98**	**0,46**	0,95	**0,97**	**0,72**	**0,69**

precision and recall measures, is commonly used in the information retrieval area as performance measure.

The $G-mean$ results on all datasets confirmed the good performances of our approach $SRS-NDF$. As we can see, our method gives better results comparing to $RS-NDF$. The Gmean of accuracies, measured separatly on each class, is associated to a point in theROC curve and the idea is to maximize the accuracies of both classes while keeping them blanced.

5 Conclusion

This paper introduced a filter ensemble selection method to improve the Random Subspace Novelty Detection Filter ($RS-NDF$) by adaptively trading off diversity and accuracy according to the data. The proposed approach $SRS-NDF$ is based on the orthogonal projection operators, the bootstrap method and the ensemble selection paradigm. Several metrics are computed on publicly available datasets and significant improvements were obtained by $SRS-NDF$ comparing to existing methods.

References

1. Markou, M., Singh, S.: Novelty detection: a review - part 1: statistical approaches. Signal Processing 83, 2481–2497 (2003)
2. Markou, M., Singh, S.: Novelty detection: a review - part 2: neural network based approaches. Signal Processing 83, 2499–2521 (2003)
3. Kohonen, T., Oja, E.: Fast Adaptive Formation of Orthogonalizing Filters and Associative Memory in Recurrent Networks of Neuron-Like Elements. Biological Cybernetics 21, 85–95 (1976)
4. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
5. Freund, Y., Saphire, R.E.: Experiments with a new boosting algorithm. In: The 13th International Conference on Machine Learning, pp. 276–280 (1996)
6. Kassab, R., Lamirel, J.-C., Nauer, E.: Novelty Detection for Modeling Users Profile. In: The 18th International FLAIRS Conference, pp. 830–831 (2005)
7. Kassab, R., Alexandre, F.: Incremental Data-driven Learning of a Novelty Detection Model for One-Class Classification Problem with Application to High-Dimensional Noisy Data. Machine Learning 74(2), 191–234 (2009)
8. Breiman, L.: Random forest. Machine Learning (2001)
9. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, Irvine (2007)
10. Bradeley, P.W.: The use of area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159 (1997)
11. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning 30, 195–215 (1998)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representation by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructures of Cognition, pp. 318–362. MIT Press (1986)
13. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics. Springer, Berlin (1986)
14. Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation (1999)
15. Greville, T.N.E.: Some applications of the pseudoinverse of a matrix. SIAM Rev. (1960)
16. Zhang, Y., Burer, S., Street, W.N.: Ensemble prunning via semi-denite programming. Journal of Machin Learning Reasearch 7, 1315–1338 (2006)
17. Hamdi, F., Bennani, Y.: Learning Random Subspace Detection Filter. In: International Joint Conference in Neural Networks, IJCNN (2011)
18. Caruna, R., Niculescu Mizil, A., Grew, G., Ksikes, A.: Ensemble selection from librairiesof models. In: The 21st International Conference on Machin Learning (2004)
19. Catell, R.: The scree test for the number of factor. Multivariate Behaviorial Research, 245–276 (1966)
20. García, V., Mollineda, R.A., Sánchez, J.S.: Index of balanced accuracy: A performance measure for skewed class distributions. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 441–448. Springer, Heidelberg (2009)