Incremental Learning and Forgetting in One-Class Classifiers for Data Streams

Bartosz Krawczyk and Michał Woźniak

Abstract. One-class classification and novelty detection is an important task in processing data streams. Standard algorithms used for this task cannot efficiently handle the changing environment to which they are applied. In this paper we present a modification of Weighted One-Class Support Vector Machine that is able to swiftly adapt to changes in data. This was achieved by extending this classifier by adding incremental learning and forgetting procedures. Both addition of new incoming data and removal of outdated objects is carried out on the basis of modifying weights assigned to each observation. We propose two methods for assigning weights to incoming data and two methods for removing the old objects. These approaches work gradually, therefore preserving useful characteristic of the examined dataset from previous iterations. Our approach was tested on two real-life dynamic datasets and the results prove the quality of our proposal.

Keywords: machine learning, one-class classification, data streams, concept drift, incremental learning, forgetting.

1 Introduction

One-class classification (OCC) is one of the most challenging areas of machine learning.It is assumed that during the classifier training stage we have at our disposal objects coming from a single class distribution, referred to as the target concept. During the exploitation phase of such a classifier there may appear new objects, that were unseen during the classifier building step. They are known as outliers. Therefore one-class classification aims at deriving a classification boundary that may separate the known target objects from possible outliers that may appear.

Bartosz Krawczyk · Michał Woźniak
Department of Systems and Computer Networks, Wroclaw University of Technology,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
e-mail: {bartosz.krawczyk,michal.wozniak}@pwr.wroc.pl

R. Burduk et al. (Eds.): *CORES 2013*, AISC 226, pp. 319–328.
DOI: 10.1007/978-3-319-00969-8_31

No assumptions about the nature of outliers should be made. The term single-class classification originates from [10], but also outlier detection or novelty detection [4] are used to name this field of study.

This is a difficult task that leads to many open problems. How the boundary should be tuned - if it is too general many unwanted outliers would be accepted, if it is to fitted to the training set then a strong overfitting may occur. Therefore it is risky to rely only on a single given model and in recent years there have been several successful attempts on how to improve the quality of one-class recognition systems [11, 12].

In the beginning the data streams originated in the financial markets. Today, the data streams are to be found everywhere - in the Internet, monitoring systems, sensor networks and other domain [9]. The data stream is very different from the traditional static data. It is an infinite amount of data that continuously arrives and the processing time is of a crucial value. Mining data streams poses many new challenges [1].

Most of the existing work on OCC has not explicitly dealt with the changing nature of the input data [14]. They are based on an underlying assumption that the training data set does not contain any uncertainty information and properly represents the examined concept. However, data in many real-world applications change their nature over time - which is a problem frequent for data streams [8]. For example, in environmental monitoring applications data may change according to the examined conditions and what once was considered an outlier may in near future become a representative of the target concept. This kind of dynamic information, typically ignored in most of the existing one-class learning methods, is critically important for capturing the full picture of the examined phenomenon. Therefore there is a need for introducing novel, adaptive techniques for dealing with non-stationary data sets [7].

In this paper we present a novel adaptive Weighted One-Class Support Vector Machine that is able to change itself according to the nature of received data streams. We propose to use the principles of incremental learning and forgetting to allow the changes in the shape of the decision boundary for the new chunks of data. The learning and forgetting in data streams is realized by modifying weights assigned to objects - we propose how to calculate weighs for new incoming objects in order to use their information to change the classifier and how to forget the old objects to prevent the overfitting of the classifier and uncontrolled increase in the volume of the dataset.

2 Weighted One-Class Support Vector Machine

One-class classification aims at distinguishing between the available objects coming from the target distribution ω_T and unknown outliers ω_O, that are unavailable during the classifier training step but may appear in the process of classifier exploitation. One-Class Support Vector Machine (OCSVM) [16] achieves this goal by computing

a closed boundary in a form of a hypersphere enclosing all the objects from ω_T. During the exploitation phase a decision made about the new object is based upon checking whether it falls inside the hypersphere. If so, the new object is labeled as one belonging to ω_T. Otherwise it belongs to ω_O.

The center a and a radius R are the two parameters that are sufficient for describing such a decision hypersphere. To have a low acceptance of the possible outliers the volume of this d-dimensional hypersphere, which is proportional to R^d, should be minimized in such a way that tightly encompasses all available objects from ω_T. The minimization of R^d implies minimization with respect to R^2. Following this the minimization functional may be formulated as follows:

$$\Theta(a,R) = R^2, \tag{1}$$

with respect to the constraint:

$$\forall_{1\leq i\leq N}: \quad \|x_i - a\|^2 \leq R^2, \tag{2}$$

where x_i are objects from ω_T, and, N stands for the quantity of training objects. Additionally to allow the fact that there may have been some outliers in the training set and to increase the robustness of the trained classifier some objects with distance to a greater than R are allowed in the training set, but associated with an additional penalty factor. This is done identically as in a standard SVM by the introduction of slack variables ξ_i.

This concept can be further extended to a Weighted One-Class Support Vector Machine (WOCSVM) [3] by the introduction of weighrs w_i that allows for an association of an importance measure to each of the training objects. This forces slack variables ξ_i, to be additionally controlled by w_i. If with object x_i there is associated a small weight w_i then the corresponding slack variable ξ_i indicates a small penalty. In effect, the corresponding slack variable will be larger, allowing x_i to lie further from the center a of the hypersphere. This reduces an impact of x_i on the shape of a decision boundary of WOCSVM.

By using the above mentioned ideas we can modify the minimization functional:

$$\Theta(a,R) = R^2 + C\sum_{i=1}^{N} w_i\xi_i, \tag{3}$$

with the modified constraints that almost all objects are within the hypersphere:

$$\forall_{1\leq i\leq N}: \quad \|x_i - a\|^2 \leq R^2 + \xi_i, \tag{4}$$

where $\xi_i \geq 0$, $0 \leq w_i \leq 1$. Here C stands for a parameter that controls the optimization process - the larger C, the less outliers are allowed with the increase of the volume of the hypersphere.

For establishing weights we may use techniques dedicated to a weighted multiclass support vector machines [5]. In this paper we propose to use a following formula:

$$w_i = \frac{|x_i - x_{mean}|}{R + \delta}, \tag{5}$$

where $\delta > 0$ is used to prevent the case of $w_i = 0$.

3 Incremental Learning and Forgetting in One-Class Classification

We assume that the classified data stream is given in a form of data chunks. At the beginning we have at our disposal an initial dataset $\mathscr{DS}_0$ that allows for training the first phase of classifier. Then with each k-th iteration we receive an additional chunk of data labeled as $\mathscr{DS}_k$. We assume that there is a possibility of concept drift presence in the incoming chunks of data. Therefore it would be valuable to adjust our one-class classifier to the changes in the nature of data.

In case when we will be using a WOCSVM trained on $\mathscr{DS}_0$ for all new incoming data we will notice a significant drop in performance - and after few new chunks of data it is possible that the performance of our model will drop below the quality of a random classifier. To prevent this from happening we propose to adapt incrementally the one-class classifier with the new incoming data to deal with the presence of concept drift and allow for a more efficient novelty detection in data streams.

We propose to apply the classifier adaptation in a changing environment via modification of weights assigned to objects from the dataset. We propose the incremental learning procedure, meaning that the dataset $\mathscr{DS}$ will consist of all available chunks of data at given k-th moment.

3.1 Incremental Learning

We propose to extend the WOCSVM concept by adding an incremental learning principle to it [15]. We use passive incremental learning. In this method new data are added without considering the importance of data. Namely, all new data are added to the training dataset.

As we associate the process of incremental learning with weights assigned to objects we propose to modify the original decision boundary by two methods for calculating weights for objects coming from a new data chunk $\mathscr{DS}_k$:

- assigning weights to objects from $\mathscr{DS}_k$ according to eq. (5). This is motivated by the fact that in the incoming data chunk not all objects should have the same impact on the shape of a new decision boundary - there may be outliers or redundant objects present.
- assigning highest weights to objects coming from the new data chunk:

$$\forall_{x_i \in \mathscr{DS}_k} : w_i = 1. \tag{6}$$

This is motivated by the fact that in the presence of the concept drift objects from a new chunk of data represent the current state of the analyzed dynamic environment and therefore should have a top priority in forming the new decision boundary.

3.2 *Incremental Forgetting*

If we apply only the incremental learning principle to the WOCSVM, the decision boundary will become more and more complex with each additional chunk of data, that enlarges our data set. This leads to a poor generalization ability in general. This can be avoided by forgetting unnecessary, outdated data [6]. It seems natural that the degree of importance of data reduces as the time passes. We propose to incorporate this concept into our concept by further expanding the WOCSVM with the incremental forgetting principle.

The simplest way is a removal of objects coming from the previous (or oldest iteration). Yet in this way we discard all the information they carried - while they still may have some valuable influence on the classification boundary (e.g., in case of a gradual concept drift where the changes of the data distribution are not rapid). A method that allows for a gradual decrease of the object influence over time seems a far more attractive idea.

Identically as in incremental learning we modify the weights to change the influence of the data on the shape of the decision boundary. In this case we propose to reduce the weights of objects from previous chunks of data in each iteration. After some given number of iterations weights assigned to them should be equal to 0 - meaning that the examined object has no longer any influence on the WOCSVM and can be safely removed from the dataset.

We propose two methods for calculating new weights for objects coming from previous iterations:

- gradual decrease of weights with the respect to their initial importance - here we introduce a denomination factor τ that is a user-specified value by which the weights will be decreased in each iteration:

$$w_i^{k+1} = w_i^k - \tau. \tag{7}$$

  This is motivated by the fact that if an object had initially assigned a higher weight it had a bigger importance for the classifier. As such these objects can be valuable for a longer period of time than objects with initial low weights - in this approach their weights will sooner approach the 0 value and they will be removed in a fewer iterations than objects with high initial value of weights.
- aligned decrease of weights without considering their initial importance - here we introduce a time factor κ that is a user-specified value standing for a number of iterations after which the object should be removed:

$$w_i^{k+1} = w_i^k - (w_i^a / \kappa), \tag{8}$$

where w_i^a stands for the initial value of the weight assigned to i-th object. As we can see the weights of objects are reduced with each iteration till they are equal to 0 (and removed from $\mathscr{DS}$) - the main difference is that this method does not consider the initial importance of data. This means that all the objects from k-th data chunk will be removed in the same moment, after κ iterations. This is motivated by the fact that changes in the dynamic environment can be unpredictable and quickly move from the original distribution - therefore data from previous steps may quickly loose its importance.

4 Experimental Investigations

4.1 Set-Up

The aims of the experiment were to assess if embedding an incremental learning and forgetting in a one-class classifier by modifying weights allows to handle the changing data streams and to compare the effectiveness of different learning and forgetting techniques introduced in this paper.

We have used two real-life datasets for the experiment:

- ECUE spam database [18] - the dataset is a collection of spam and legitimate consists of emails received by one individual. Apart from the initial training dataset there are available 12 data chunks with the presence of a gradual concept drift, consisting of messages collected over one year - single month collection in a single data chunk.
- Ozon level detection database - this data describes local ozone peak prediction, that is based on eight hours measurement with the presence of a gradual concept drift.

Each dataset was prepared to consist of 5 parts - one for initial training and five data chunks for testing the incremental learning and forgetting procedures.

For the experiment a WOCSVM with a RBF kernel and Canberr distance metric [13] is used as a base classifier. We have examined the performance of five different WOCSVM models for data stream classification:

1. L_0F_0 - an WOCSVM trained without modifying the values of weights. Here objects from new dataset are added to the training set and the forgetting was implemented as a passive forgetting - objects are removed from the dataset after $\kappa iterations$. This is a baseline model for further comparison.
2. L_1F_1 - a WCOSVM with the incremental learning by assigning weights to new objects according to eq. (5) and with forgetting by the gradual decrease of weights.
3. L_2F_1 - a WCOSVM with the incremental learning by assigning highest weights to new objects and with forgetting by the gradual decrease of weights.

4. L_1F_2 - a WCOSVM with the incremental learning by assigning weights to new objects according to eq. (5) and with forgetting by the aligned decrease of weights without considering their initial importance.
5. L_2F_2 - a WCOSVM with the incremental learning by assigning highest weights to new objects and with forgetting by the aligned decrease of weights without considering their initial importance.

The value of τ parameter was set to 0.2 and κ to 2. These values were derived using a grid-search procedure, as they offered the best performance.

The combined 5x2 cv F test [2] was carried out to asses the statistical significance of obtained results.

All experiments were carried out in the R environment [17].

4.2 Results

Results for the ECUE dataset are given in Tab. 1, while for the Ozon dataset in Tab. 2.

Table 1 Results of the experiment with the respect to the accuracy [%] and statistical significance for ECUE dataset. Small numbers under each method stands for the indexes of models which the considered one outperforms (in a statistically significant way).

Method	$\mathscr{DS}_0$	$\mathscr{DS}_1$	$\mathscr{DS}_2$	$\mathscr{DS}_3$	$\mathscr{DS}_4$
$L_0F_0^1$	84.56	76.34	73.25	73.03	72.44
	–	–	–	–	–
$L_1F_1^2$	84.56	80.34	79.65	79.22	77.43
	–	1,4	1,4	1,4,5	1
$L_2F_1^3$	84.56	82.11	81.03	80.04	79.05
	–	*ALL*	*ALL*	1,4,5	1,2
$L_1F_2^4$	84.56	78.96	78.23	78.05	78.20
	–	1	1	1	1,2
$L_1F_2^5$	84.56	81.54	80.15	78.71	78.55
	–	1,2,4	1,4	1	1,2

4.3 Results Discussion

From the experimental results one may see that proposed methods for incremental learning and forgetting for WOCSVM are a valuable tool for dealing with the changing environment in data streams. The standard OCSVM with simple method for adapting to new data was unable to deal with gradual concept drift present in the examined datasets.

Four tested models based on two learning and two forgetting procedures outperformed this baseline solution on all data. Modifying the weights assigned to objects instead of simply adding them and removing allowed for a smoother change of

Table 2 Results of the experiment with the respect to the accuracy [%] and statistical significance for Ozone dataset. Small numbers under each method stands for the indexes of models which the considered one outperforms (in a statistically significant way).

Method	$\mathscr{DS}_0$	$\mathscr{DS}_1$	$\mathscr{DS}_2$	$\mathscr{DS}_3$	$\mathscr{DS}_4$
$L_0F_0^1$	87.44	84.05	82.90	79.45	76.48
	–	–	–	–	–
$L_1F_1^2$	87.44	85.92	84.11	81.68	80.04
	–	1	1	1,5	1
$L_2F_1^3$	87.44	86.25	85.55	82.65	83.11
	–	*ALL*	*ALL*	1,2,5	*ALL*
$L_1F_2^4$	87.44	85.22	84.86	82.70	80.15
	–	1	1,2	1,2,5	1
$L_1F_2^5$	87.44	85.80	84.89	80.97	81.98
	–	1	1,2	1	1,2,4

model and introduced elasticity into OCC stream data classification. Weight manipulation lead to a situation in which the data from previous chunks had neither too big or to small influence on the shape of the decision boundary. Therefore this approach preserved the valuable influence of the older data, while adapting to changes in the incoming objects.

Out of four tested combination the most promising results were returned by the model based on the incremental learning by assigning highest weights to new objects and the forgetting by the gradual decrease of weights. In most cases it statistically outperformed all other models. This means that during the concept drift the new objects should have the biggest influence on the shape of the new boundary. As for the forgetting procedure the experiments favored the approach in which relevant objects (i.e., with high initial weights) influence the decision boundary for larger number of iterations than those with low initial weights.

Interestingly there was no stable trend in differences between all other methods, which leads to a conclusion that their performance is data set-related and needs further experimental analysis.

The model response to the presence of the concept drift may display itself by the change of the center or/and radius of the hypersphere. In both experiments the proposed methods responded by a continuous shift of the sphere center towards the direction of the drift. In comparison the radius of the hypersphere was subject to much lower variance.

5 Conclusions

One-class classification and novelty detection in data streams is a promising research direction. There is an ongoing need for introducing classifier models that

can adapt to changing environment. In this paper we have introduced a modified Weighted One-Class Support Vector Machine, augmented with the principles of incremental learning and forgetting. These techniques allowed to adapt the shape of the decision boundary to changes in the stream of data.

We proposed to adapt WOCSVM by modifying weights that are assigned to objects in the set. We have applied the incremental learning by two approaches based on calculating weights for incoming objects. Incremental forgetting was applied to avoid constructing too complex model and to reduce the volume of the dataset, which is an important problem in distributed applications. Forgetting was applied as decreasing the weights assigned to objects over time, to a point in which they no longer influence the WOCSVM and may be removed from the data set.

In our future works we would like to test our approach on different types of concept drift, improve the incremental learning and forgetting procedure by making it selective (i.e., choosing only some part of the data to add to the set or remove from it) and combining our incremental WOCSVM in ensembles.

Acknowledgements. This work is supported by the Polish National Science Center under a grant N N519 650440 for the period 2011-2014.

References

1. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: On demand classification of data streams. In: KDD 2004 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 503–508 (2004)
2. Alpaydin, E.: Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. Neural Computation 11(8), 1885–1892 (1999)
3. Bicego, M., Figueiredo, M.A.T.: Soft clustering using weighted one-class support vector machines. Pattern Recognition 42(1), 27–32 (2009)
4. Bishop, C.M.: Novelty detection and neural network validation. IEE Proceedings: Vision, Image and Signal Processing 141(4), 217–222 (1994)
5. Cyganek, B.: One-class support vector ensembles for image segmentation and classification. Journal of Mathematical Imaging and Vision 42(2-3), 103–117 (2012)
6. Gent, I.P., Miguel, I., Moore, N.C.A.: An empirical study of learning and forgetting constraints. AI Communications 25(2), 191–208 (2012)
7. Gomez-Verdejo, V., Arenas-Garcia, J., Lazaro-Gredilla, M., Navia-Vazquez, A.: Adaptive one-class support vector machine. IEEE Transactions on Signal Processing 59(6), 2975–2981 (2011)
8. Hashemi, S., Yang, Y., Mirzamomen, Z., Kangavari, M.: Adapted one-versus-all decision trees for data stream classification. IEEE Transactions on Knowledge and Data Engineering 21(5), 624–637 (2009)
9. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 97–106 (2001)
10. Koch, M.W., Moya, M.M., Hostetler, L.D., Fogler, R.J.: Cueing, feature discovery, and one-class learning for synthetic aperture radar automatic target recognition. Neural Networks 8(7-8), 1081–1102 (1995)

11. Krawczyk, B.: Diversity in ensembles for one-class classification. In: Pechenizkiy, M., Wojciechowski, M. (eds.) New Trends in Databases & Inform. AISC, vol. 185, pp. 119–129. Springer, Heidelberg (2012)
12. Krawczyk, B., Woźniak, M.: Combining diverse one-class classifiers. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS, vol. 7209, pp. 590–601. Springer, Heidelberg (2012)
13. Krawczyk, B., Woźniak, M.: Experiments on distance measures for combining one-class classifiers. In: Proceedings of the FEDCISIS 2012 Conference, pp. 88–92 (2012)
14. Masud, M., Gao, J., Khan, L., Han, J., Thuraisingham, B.M.: Classification and novel class detection in concept-drifting data streams under time constraints. IEEE Transactions on Knowledge and Data Engineering 23(6), 859–874 (2011)
15. Ross, D.A., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. International Journal of Computer Vision 77(1-3), 125–141 (2008)
16. Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. In: Adaptive Computation and Machine Learning. MIT Press (2002)
17. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008)
18. Zmyslony, M., Krawczyk, B., Woźniak, M.: Combined classifiers with neural fuser for spam detection. In: Herrero, A., Snasel, V., Abraham, A., Zelinka, I., Baruque, B., Quintin, H., Calvo, J.L., Sedano, J., Corchado, E. (eds.) International Joint Conference CISIS12-ICEUTE12-SOCO12 Special Sessions. AISC, vol. 189, pp. 245–252. Springer, Heidelberg (2012)