

Detecting and Learning from Unknown by Extremely Weak Supervision: eXploratory Classifier (xClass)

Plamen Angelov · Eduardo Soares*

Received: date / Accepted: date

Abstract In this paper, we break with the traditional approach to classification, which is regarded as a form of supervised learning. **We offer a method and algorithm, which make possible fully autonomous (unsupervised) detection of new classes, and learning following a very parsimonious training priming (few labeled data samples only).** Moreover, new unknown classes may appear at a later stage and the proposed xClass method and algorithm are able to successfully discover this and learn from the data autonomously. Furthermore, the features (inputs to the classifier) are automatically sub-selected by the algorithm based on the accumulated data density per feature per class. In addition, the automatically generated model is easy to interpret, is locally generative and based on prototypes which define the modes of the data distribution. As a result, a highly efficient, lean, human-understandable, autonomously self-learning model (which only needs an extremely parsimonious priming) emerges from the data. To validate our proposal we approbated it on four challenging problems, including imbalanced Faces-1999 data base, Caltech-101 data set, vehicles dataset, and iRoads dataset, which is a dataset of images of autonomous driving scenarios. Not only we achieved higher precision (in one of the problems outperforming by 25% all other methods), but, more significantly, we only used a single class beforehand, while other methods used all the available classes) and we generated interpretable models with smaller number of features used, through extremely weak and weak supervision. We demonstrated the ability to detect and learn new classes for both, images and numerical examples.

Plamen Angelov
School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK
E-mail: p.angelov@lancaster.ac.uk

Eduardo Soares
School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK
E-mail: e.almeidasoares@lancaster.ac.uk

**Corresponding Author*

Keywords Extremely weak supervision · Novelty detection · Interpretability · Autonomously self-learning model

1 Introduction

Machine learning and pattern recognition, including classification are perhaps at the peak of their development with a sharp interest not only from scientists and practitioners, but also from the wider public and media. This is, in part, thanks to the boom surrounding the wider area of artificial intelligence (AI) and recent successful and widely publicized applications ranging from games [34, 14], driverless cars [33, 10], defense and security [1, 32, 35], home applications [23, 28]. Despite the great success of the standard bearer algorithm in this area, the so called deep learning in image and speech recognition [18, 27], the underlying concept of machine learning which requires large amount of labeled training data remains unchanged. So called reinforcement learning offers some departure from complete labeling, but still requires user input for each individual data sample. The most powerful approaches such as deep learning and **support vector machines (SVM)** suffer from lack of interpretability [25, 5, 30, 11], are extremely power-, time- and computational- resources hungry and are like dinosaurs – unable to adapt and change with agility. They require complete retraining even for a single or few new data samples.

In this paper we propose a method and algorithm that departs from the traditional approach and offers a paradigm shift bringing the machine learning, in general, and pattern recognition and classification, in particular, extremely close to a fully unsupervised form. In a nutshell, it offers a self-learning locally generative models that work together and require extremely light supervision in the form of few data samples. It is able to automatically detect the unknown and to learn from it. This is in sharp contrast to the traditional approach where learning is, in essence, only an averaging of the history. The current approaches struggle to detect changes, dynamical evolution or appearance of new classes. They also assume a certain number of features (the same for all classes) provided at the start of the process. This is one of the reasons traditional approaches struggle to predict or react quickly to sudden changes in the data pattern, such as the economic crash during 2008 [15], for example.

Methods like eClass [8], FLEXFISClass [20] and other similar ones are called “evolving” classifiers. They are designed to take into account new coming data samples. However, when talking about new classes (rather than just new data samples) class label is required which means these methods are supervised learning methods. The proposed method in this paper is unsupervised in regards to the new data that represent a new class. There are also unsupervised evolving algorithms for clustering [9], but these methods do not deal with classification as the method proposed in this paper. Another type of methods that claim to approach similar problems are the so-called zero-shot learning (ZSL) methods. They have as an objective to transfer a learnt model to unknown classes with-out the acquisition of new features. However, the main

problem with this type of technique is the dependence on additional information to relate unknown classes to previously trained models. Not always such information is available or possible to acquire [17]. In this respect, the ZSL approach is not unsupervised in terms of the new class and not a direct comparator.

The proposed approach is prototype-based and learns locally around them extracting the empirical data distribution called typicality as well as the data density [6]. The approach is recursive, thus computationally very lean. It is also non-iterative, non-parametric. This adds to its efficiency in terms of time and computational resources. From the user perspective, the proposed approach is clearly understandable to human users since it can be represented in a linguistic *IF...THEN* form. It combines reasoning and logic with machine learning. It can also be presented as a deep neural network. Finally, it also has a statistical nature and offers an empirical form of the **probability density function (pdf)** [7].

In this paper we apply this new principally different type of machine learning to four challenging problems and demonstrate its significant advantages. **The main challenges that the method proposed in this paper addresses are: i) to detect when a certain unlabeled (new) data sample is not from a class that was used in training, i.e. to have class "Unknown" or "New"; ii) to learn from such new unlabeled data in an unsupervised manner.** The proposed approach to address the first issue is based on the drop of the density that represent the confidence in a decision. The proposed approach to the second issue is by learning from the data for which the class is "New". The proposed approach further selects prototypes out of the data samples of the "New" class according to their density in the same way as for the other/known classes. Because, the learning in the proposed approach is per class, all new data from a "New" class are analysed separately from the data from the known classes. The remainder of this paper is organized as follows: The method and algorithm section introduces the proposed exploratory approach for extremely weakly supervised classification. The experimental data employed in the analysis and results are presented in the results section. Discussion is presented in the last section of this paper.

2 Material and Methods

2.1 Concept and Basic Algorithm

Traditionally, the pipeline of learning from data includes the following steps:

- 1) **Pre-processing**, which includes different substeps like normalization/standardization, dealing with missing data, and feature selection [16]. Specifically for image processing there are often other stages, such as rotation, augmentation, scaling, elastic deformation, etc [26]. Even deep learning methods which claims to avoid handcrafting apply some of the cited steps.

2)**Learning phase**, which can be offline, when the full dataset is available; or it can be done online, when the data arrive in the form of a data stream (sample-by-sample). Evolving learning, ability of the algorithms to adapt their parameters and structure according to the non-stationary data streams, is a more sophisticated form of online learning [3, 29].

3)**Generating outputs** for new unseen data, which is the **validation phase**. Different algorithms use different strategies in order to validate the model generated in the learning phase.

The proposed method also starts with a pre-processing step which involves mostly the same steps depending on the specific problem. For example, for image processing we may also apply scaling, augmentation, rotation, etc. Practically for all problems normalization and standardization is required.

The proposed xClass method uses standardization and normalization as follows:

Firstly, it standardize the newly observed data sample, x_i ; where $i = 1, 2, \dots, n$ denotes a time stamp in the current moment. $j = 1, 2, \dots, n$ refers to the number of features of the given x .

$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu(x_{i,j})}{\sigma(x_{i,j})} \quad (1)$$

where \hat{x} denotes the standardized data sample. Outliers ($|\hat{x}| \geq 3$) are ignored and not used for training. After that, the data is rescaled within the range $[0, 1]$ to consider them in the same proportion. It is important to highlight that in the proposed xClass method, the normalization is done upon the standardized data. Unity-based normalization of the i -th element of the j -th sample is given by:

$$\bar{x}_{i,j} = \frac{\hat{x}_{i,j} - \min_i(\hat{x}_{i,j})}{\max_i(\hat{x}_{i,j}) - \min_i(\hat{x}_{i,j})} \quad (2)$$

where \bar{x} denotes the normalized data sample.

The prototype-based learning is the core of the proposed method which represents local (the prototypes are focal points of locally valid generative models described by multimodal Cauchy distribution [6]. The meta-parameters are initialized with the first observed data sample. The proposed algorithm works per class; therefore, all the calculations are done for each class separately.

$$P \leftarrow 1; \quad \mu \leftarrow \bar{x}_i; \quad (3)$$

where μ denotes the global mean of data samples of the given class. P is the number of the identified prototypes in total from the observed data samples.

Each class C is initialized by the first data sample of that class:

$$\begin{aligned} C_1 &\leftarrow \{\bar{x}_1\}; & p_1 &\leftarrow \bar{x}_1; \\ S_1 &\leftarrow 1; & r_1 &\leftarrow r^*; \end{aligned} \quad (4)$$

where, p_1 is the prototype of C_1 ; S_1 is the corresponding support (number of members); r_1 is the corresponding radius of the area of influence of C_1 .

In this paper, we use $r^* = \sqrt{2 - 2\cos(30^\circ)}$ same as [6]; the rationale is that two vectors for which the angle between them is less than $\pi/6$ or 30° are pointing in close/similar directions. That is, we consider that two feature vectors can be considered to be similar if the angle between them is smaller than 30 degrees. Note that r^* is data derived, not a problem- or user- specific parameter. In fact, it can be defined without *prior* knowledge of the specific problem or data. The next step is to calculate the data density at \bar{x}_i and p_j ($j = 1, 2, \dots, P$).

$$D(\bar{x}_i) = \frac{1}{1 + \frac{\|\bar{x}_i - p_j\|^2}{(\sigma_i)^2}} \quad (5)$$

where p_j ($j = 1, 2, \dots, P$) is the set of prototypes, and σ_i is the standard deviation.

The reason it is Cauchy is not arbitrary [4]. It can be demonstrated theoretically that if Euclidean or Mahalanobis type of distances in the feature space are considered, the data density reduces to Cauchy type as referred in equation (5). It can also be demonstrated that the so called *typicality*, τ , which is the weighted average of the data density, D , with weights representing the frequency of occurrence of a data sample [6]. Furthermore, the *typicality*, τ can be considered an empirically derived form of the pdf having the same properties, notably, it integrates to 1 an infinite range.

Density per feature f is obtained according to the equation (5), where D_i^f denotes the density for f -th feature of the \bar{x}_i sample.

The cumulative effect across all data samples per feature can be obtained according to the equation (6).

$$\Lambda_i^f = \frac{\sum_{i=1}^n D_i^f(\bar{x}_i^f)}{n}. \quad (6)$$

The cumulative contribution for each feature Λ_i^f can be rank ordered, n represents the number of samples. The higher, the value of Λ_i^f is for a particular feature, the more important is the f -th feature. The rationale is that an interesting feature has higher density than other features - meaning that it conveys unique, different clear information, and, as a consequence, it contributes more to the classifier's result because the overlap between data of different classes is less pronounced for this feature.

Then the algorithm absorbs the new data samples one by one by assigning them to the nearest (in the feature space) prototype:

$$n^* = \underset{j=1,2,\dots,P}{\operatorname{argmin}} (\|\bar{x}_i - p_j\|^2) \quad (7)$$

Because of this form of assignment, the shape of the data partitioning is of the so-called Voronoi tessellation type [21]. We call all data points associated with a prototype *data clouds*, because their shape is not regular (e.g., hyper-spherical, hyper-ellipsoidal, etc.) and the prototype is not necessarily the statistical and geometric mean [6].

In case, the following condition [6] is met:

$$\begin{aligned} IF (D_i(\bar{x}_i) \geq \max_{j=1,2,\dots,P} D_i(p_j)) \quad OR \quad (D_i(\bar{x}_i) \leq \min_{j=1,2,\dots,P} D_i(p_j)) \\ THEN (add a new data cloud) \end{aligned} \quad (8)$$

It means that \bar{x}_i is out of the influence area of p_j . Therefore, \bar{x}_i becomes a new prototype of a new *data cloud* with meta-parameters initialized by equation (9). Add a new data cloud:

$$P \leftarrow P + 1; \quad C_P \leftarrow \{\bar{x}_i\}; p_P \leftarrow \bar{x}_i; \quad S_P \leftarrow 1; \quad r_P \leftarrow r_o; \quad (9)$$

Otherwise, data cloud parameters are updated online by equation (10). It has to be stressed that all calculations per data cloud are performed on the basis of data points associated with a certain data cloud only (i. e. locally, not globally, on the basis of all data points).

$$\begin{aligned} C_{n^*} \leftarrow C_{n^*} + \{\bar{x}_i\}; \quad p_{n^*} \leftarrow \frac{S_{n^*}}{S_{n^*} + 1} p_{n^*} + \frac{S_{n^*}}{S_{n^*} + 1} \bar{x}_i; \\ S_{n^*} \leftarrow S_{n^*} + 1; \quad r_{n^*}^2 \leftarrow \frac{r_{n^*}^2 + (1 - \|p_{n^*}\|^2)}{2}; \end{aligned} \quad (10)$$

One of the strongest aspects of the proposed approach is its high level of interpretability which comes from its prototype-based, local generative models as well as as its ability to be expressed as a set of linguistic *IF...THEN* fuzzy rules of the following type:

$$R: \quad IF (x \sim p_1) \quad OR \quad \dots \quad OR \quad (x \sim p_P) \quad THEN (Class \ c) \quad (11)$$

The fuzziness represents the degree of association/similarity to the prototypes. Indeed, the value of data density, D , equation (5) can be interpreted as a membership function of the fuzzy set $(x \sim p)$ [6]. With a maximum 1 when $x = p$. The continuous typicality, τ given by the equation (12), is an empirically derived form of probability distribution. The value of τ even at the point $x = p_i$ is much less than 1 the integral of $\int_{-\infty}^{\infty} \tau dx = 1$.

$$\tau_i(x) = \frac{D_i(x)}{\int_x D_i(x) dx} \quad (12)$$

Detect & Learn from Unknown

This is the most innovative part of the proposed algorithm in addition to the feature selection per class, which makes it exploratory (we call it xClass) and allows to detect new data patterns autonomously and learn from them.

Drop of confidence (detect the novelty)

Unlabeled data samples become available as soon as the training process with labeled samples finishes. Then, the eXploratory classifier (xClass) can continue to learn from these unknown data samples. The unlabeled training samples are defined as the set $\{u\}$, and the number of unlabeled samples is defined as U .

The first step in the weakly supervised learning process of xClass, is to extract the vector of confidence/degrees of closeness to the nearest prototypes for each unlabeled data sample defined as $\lambda(u_i)$, $i = 1, 2, \dots, U$ follows:

$$\lambda = \max_{j=1,2,\dots,P} (\bar{x}, p_j), \quad (13)$$

where λ denotes the confidence degree.

The recursive mean $\bar{\mu}_i$ of the λ^{max} for the labeled data samples is used to detect sudden drop of the confidence generated by the xClass classifier when a new unknown class arrives and can be calculated as follows [2]:

$$\bar{\mu}_i = \frac{i-1}{i} \bar{\mu}_{i-1} + \frac{1}{i} \lambda_i^{max}, \bar{\mu}_1 = \lambda_1^{max}. \quad (14)$$

Then the $m\text{-}\sigma$ rule is applied, for detailed explanation about the $m\text{-}\sigma$ please refer to [24]. New classes are actively added by the proposed xClass classifier when the inequality (15) is satisfied and rules are actively created. Otherwise, if the inequality is not satisfied the newly arrival unlabeled data samples are used for updating the structure and meta-parameters of the xClass classifier. Fig. 1 illustrates the drop of confidence of the proposed method when a new unseen class arrives. The black line indicates the confidence of xClass. As the fall is detected, if the inequality (15) is satisfied this indicates that the label of this data sample is not any of the known to xClass labels. The options are that: a) This drop is a one off due to outlier, noise, randomness, or b) a number of such data samples above a drop of confidence is detected are close to each other in the data space (please note that they may not necessarily arrive one after the other as in Fig. 1). Otherwise, if the condition given by the inequality (15) is not met the data sample is used to update the meta-parameters of the proposed method.

$$\begin{aligned} & \text{IF } \lambda^{max}(U_i) < (\bar{\mu}_i - m\sigma) \text{ THEN } (U_i \in \text{Possible new class detected}) \\ & \text{ELSE } (\text{Update structure and meta - parameters}) \end{aligned} \quad (15)$$

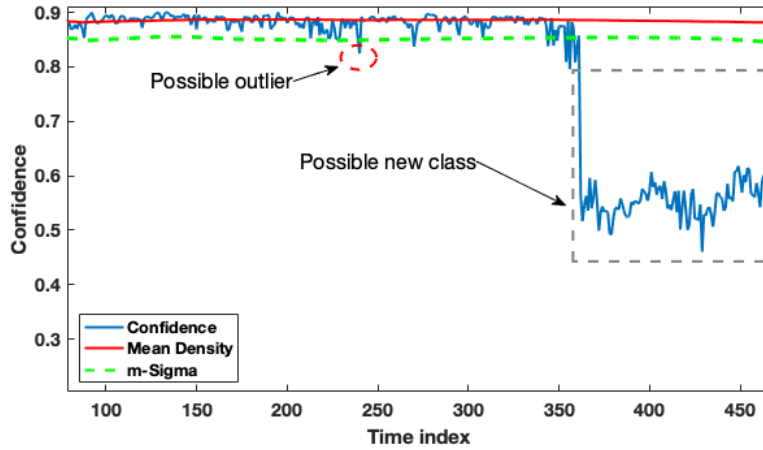


Fig. 1 Drop of confidence of the proposed method when a new a unseen class arrives

When the inequality (15) is satisfied, the arrival data sample is denoted as a potential outlier and temporally saved. When several of potential outliers are close to each other in the data space, have similar densities, they are denoted as ‘new class 1’, if more than one group is formed than new classes are formed as well and new labels as ‘new class 2’ are generated. The user can be proactively asked to (optionally) label with a semantically meaningful identification, for example, ‘apple’, however, no retraining is required.

One or few labels for new detected classes are provided. The validation process is done through the ‘winners-take-all’ principle, which is given by,

$$Label = \operatorname{argmax}(\lambda(\bar{x})). \quad (16)$$

The general structure of the proposed xClass approach is illustrated by the block diagram presented in Fig. 2.

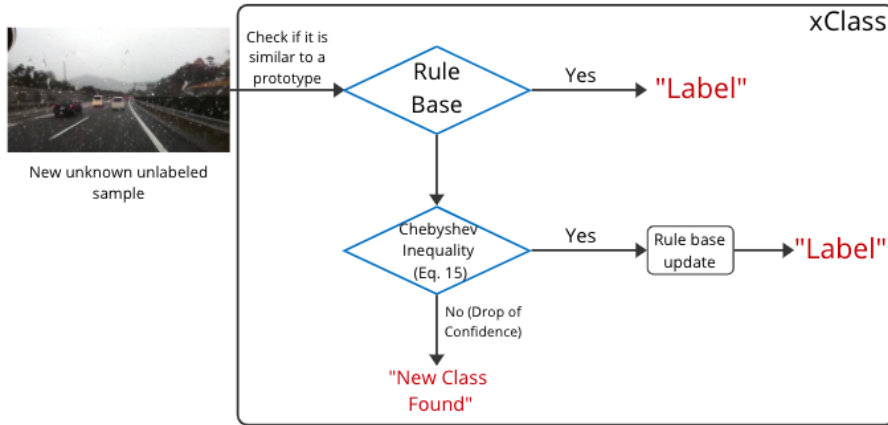


Fig. 2 General structure of xClass – block diagram

3 Results

In this section we will demonstrate the results obtained by the proposed extremely weakly supervised classification approach. Computational simulations were performed to assess the accuracy of the classification methods considering 4 different benchmark problems. The results from experimentation with the proposed algorithm aim to demonstrate that it offers:

- high precision as compared with the top state-of-the-art algorithms.
- ability to detect unseen/new data patterns autonomously and learn from them.
- ability to learn with extremely low supervision (few) labeled data samples for the newly detected classes.
- ability to autonomously select the most effective features per class.
- highly transparent interpretable model.
- no user- or problem- specific algorithmic parameter (except for feature selection which can be done by *ad hoc* decision).
- non-iterative algorithm able to learn continuously.

3.1 iRoads dataset

In the first experiment the iRoads dataset [22] was considered. The convolutional deep neural network VGG-16 was trained with 80% of the available iRoads dataset; however, images for the ‘Rainy day’ scenario were omitted of the training phase. After the training phase, ‘Rainy day’ trained images were presented to the neural network. As the VGG-16 approach was not trained for the presented situation, and it is not able to adapt its structure for the newly arrived class, it misclassified the ‘Rainy Day’ scenario with almost 90% confidence as a ‘Night’ scenario as illustrated by Fig. 3.

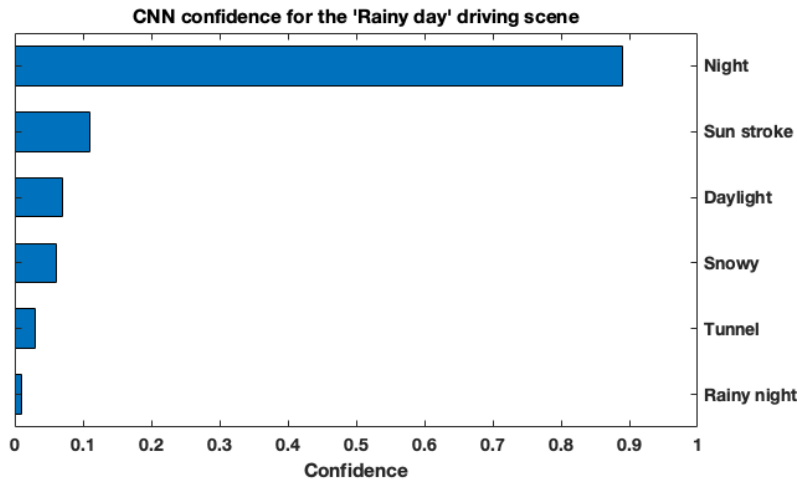


Fig. 3 Wrong classification given by VGG-16 for a new unknown class (Rainy Day).

The convolutional neural network VGG-16 misclassified with almost 90% of confidence the ‘Rainy day’ driving scenario as a ‘Night’ scenario as illustrated by Fig. 3. This is not surprising because the VGG-16 (same as other mainstream deep neural networks) can only recognise what it was trained for and is not equipped with an exploratory mechanism to enable detection and learning from unknown data samples. In such new situations mainstream deep networks require a full retraining in order to correctly classify new classes. However a full retraining of a deep neural network is usually time consuming, computational expensive, and costly and involves the human for labeling purposes.

The xClass exploratory mechanism is able to discover new classes as they arrive to the system due to its mechanism based on the recursive density estimation [2] and Chebyshev inequality approach [24] as given by Fig. 4. The blue line indicates the confidence value (λ^{max} boundary) given by the xClass classifier, the red line indicates the recursive density estimation value, the green line is the $3\text{-}\sigma$. The sudden fall of the blue line indicates the moment when the unlabeled set of images belonging to an unknown class arrive to the system.

The proposed xClass classifier was trained with 80% of the available iRoads images of all classes except the ‘Rainy day’ class. Then, the new unlabeled class was present to the proposed classifier, xClass was able to successfully detect the suddenly drastic fall in the confidence (Fig. 4) and proactively create a new class as illustrated by Fig. 5. The prototype-based and non-iterative nature of the proposed method allowed to detect the fall in the confidence (λ^{max}) in real time, and differently, from traditional deep learning approaches, no retraining is required to learn the new class.

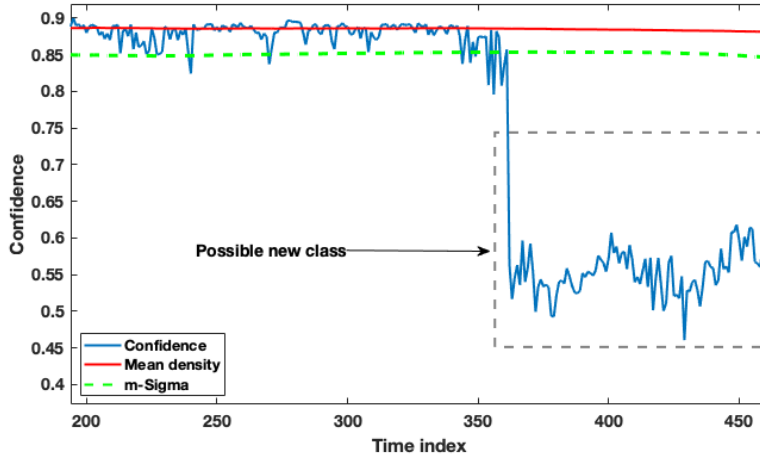


Fig. 4 Sudden drop of confidence due the presentation of new unknown classes.

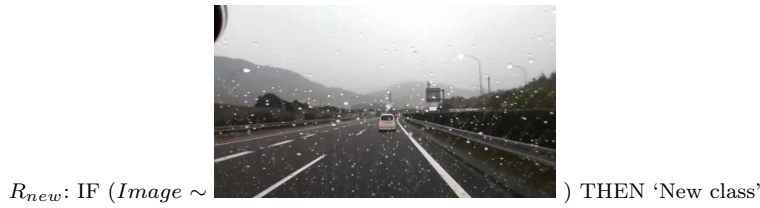


Fig. 5 A new rule is proactively created when a sudden fall in the confidence is detected through the inequality (13). The proposed xClass classifier is highly interpretable due to its rule-based nature. This advantage favors human experts analysis as it provides a transparent structure, differently from the ‘black box’ approaches such as deep neural networks.

The proposed xClass classifier obtained 99.12% classification accuracy for unlabeled images using the $3\text{-}\sigma$ approach. The semantically meaningful label ‘Rainy Day Scene’ is optional and requires only one-off involvement by the human (by default it will stay as ‘new class 1’). The final rule generated for this new class detected by the proposed xClass classifier is given by Fig. 6.

3.2 Faces-1999 dataset

As a second example, we consider the Faces-1999 dataset provided by Caltech [12]. For the faces recognition problem, the xClass classifier is trained with just one type of face, differently from traditional approaches which are primed with all available classes (20 different types of faces). We used the fully connected layer of VGG-16 for features extraction. For each image it produces 4096 values that can be considered [to be] abstract features.

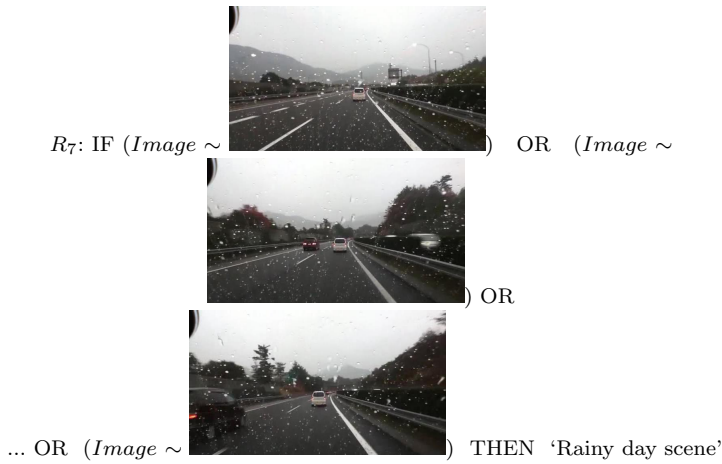


Fig. 6 Final rule given by the xClass classifier for the new detected class. Label is attached during the validation phase. Differently from ‘black box’ approaches as deep neural networks, xClass provides highly interpretable rules which can be used by human experts for different analysis as necessary.

As the traditional approaches are not equipped with exploratory mechanism, they are not able to discover new data patterns, and then, they classify new arrival data samples as the trained class. The, the proposed approach was presented to the new classes, and it was able to detect these new types of faces through the drop of confidence as illustrated by Fig. 7. After the detection of these new classes, an extremely weak supervision (1% training data labeled) and weak supervision (10% training data labeled) is provided in order to label these newly arrived. After, the labeling phase, the classification task was performed. As one can see from Fig. 8 and 9, the proposed xClass method can surpass its state-of-the-art competitors as they require more labeled data to provide good results. With just 1% of training data is clearly visible the advantage of xClass. On real scenarios the labeling process is very time consuming and is not always possible. The classification curve is given by Fig. 9.

Fig. 7 illustrates the sudden drop in the confidence when new unknown classes are presented to xClass classifier; the xClass uses the drop of confidence based on the density of the data to discover new classes. Traditional approaches are not equipped with exploratory mechanisms as the proposed xClass method; therefore, they are not able to detect new data patterns and adapt their structure to this situation. It is notable that the proposed xClass classifier can obtain better results without the necessity for huge number of labeled data, differently from traditional approaches. The performance curve is given in Fig. 9, as illustrated, with xClass still producing better classification rates when more training data is provided.

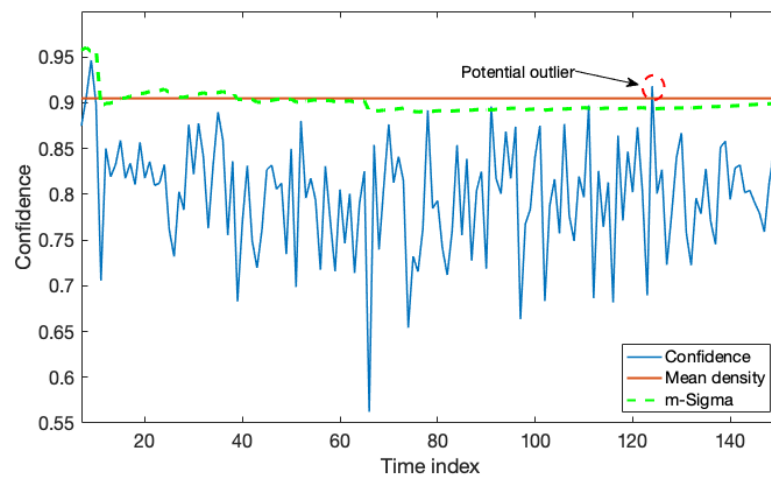


Fig. 7 Sudden drop of confidence due the presentation of new unknown classes for the Faces-1999 dataset.

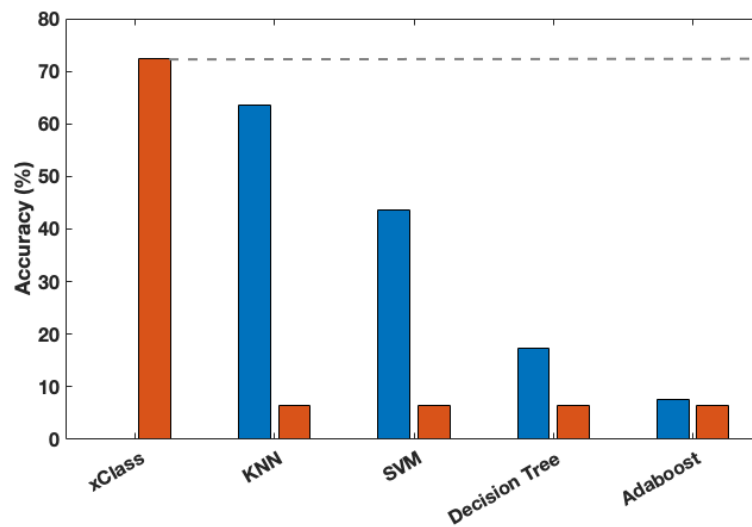


Fig. 8 Accuracy for extremely weak supervision classification for the Faces-1999 dataset. red bars illustrate the results obtained by state-of-the-art approaches when just one class is provided during the training phase. The blue bars indicate the results when all the classes are provided.

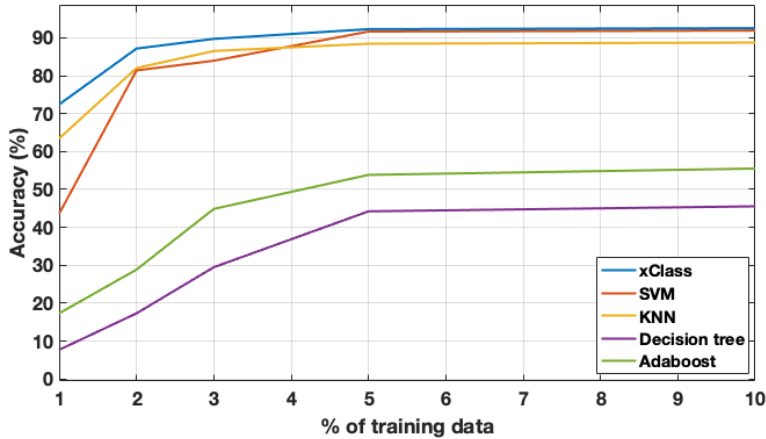


Fig. 9 Classification curve for different number of training samples for the Faces-1999 dataset.

3.3 Caltech-101 dataset

As a third case, we consider the Caltech-101 dataset [13]. As in the other experiments the proposed xClass classifier was primed with 80% of data samples from the first class for training, and then, used its exploratory mechanism to discover the other classes autonomously and learn from them based on the data density according through the drop of confidence as detailed in Fig. 10; as illustrated in Fig. 11, traditional approaches are not able to detect new data patterns after the training phase (traditional approaches were trained with just 1 class), and then, tend to produce results with low accuracy. Unlike supervised methods which are data hungry, the proposed xClass approach could obtain high classification accuracy with extremely weak supervision (Fig. 11), in order word, with less training data as possible. The acquisition of labeled data requires enormous human efforts and it is very time consuming. Fig. 12 gives the evolution of the performance of the proposed exploratory classifier as more training samples are provided. As it is illustrated by Fig. 12, the xClass classifier is able to produce better results in terms of accuracy, demonstrating its efficiency to detect and learn from unknown effectively.

The Caltech-101 dataset is constituted of 101 different classes. However, in the experiment only 10 classes were used. Supervised methods such as Decision tree, **k-nearest neighbors (KNN)**, Adaboost, and SVM require information about all the available classes beforehand, in order to produce better results (the red bars in Fig. 11 illustrate the results obtained when just one class is used in the training phase). In comparison, the proposed extremely weakly supervised approach requires just the knowledge about one class beforehand as illustrated by Fig. 10 as the other classes are discovered through its exploratory mechanism. The blue bar in Fig. 11 illustrates the classification

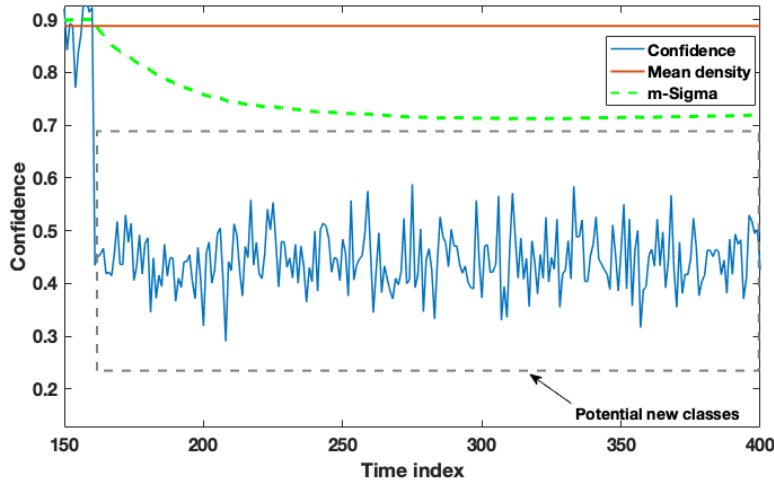


Fig. 10 Sudden drop when new unknown are classes are presented to the xClass method – Caltech-101 dataset.

results when just 1% of labeled training data is provided for all classes. The proposed exploratory xClass classifier could obtain almost 90% of classification accuracy. State-of-the-art approached have the necessity for labeled training data to produce acceptable results as illustrated in Fig. 12. Even when more labeled training data is provided, the proposed xClass classifier still produce better results in terms of accuracy than its competitors. **Furthermore, the ZSL method proposed by [19] was reported to provide 57% accuracy for the same problem which is significantly poorer result than the one obtained by the proposed xClass method. In addition to the significantly higher accuracy than the ZSL method, the proposed xClass method also has the advantage of allowing human inspection of the decision-making process (explainable).**

3.4 Vehicles dataset

In the fourth case, we consider the vehicles dataset [31], which is a non-image based dataset. xClass is, firstly, trained with just one sample of the first class, and then, it has to autonomously detect the other classes based on the empirically observed data and the sudden drop of confidence (Fig. 13). The inner parallel feature selector of the proposed approach selected 7 out of the 18 original features differently for each class. This is helpful to improve the interpretability of the proposed classifier. Results obtained by xClass and its competitors are given in Fig. 15. It is important to highlight that SVM, KNN, Decision Tree, Adaboost, **Long short-term memory (LSTM)** are all supervised methods, and they were trained with all available classes beforehand (red bars in Fig. 14 illustrate the results obtained by the traditional supervised ap-

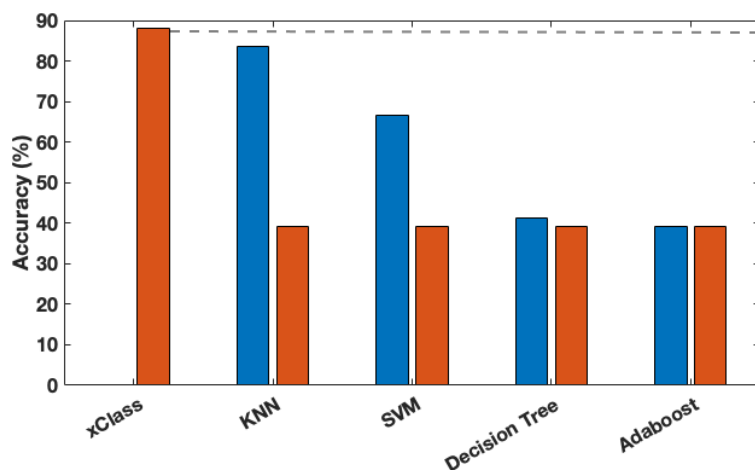


Fig. 11 Accuracy for extremely weak supervision classification for the Caltech-101 dataset.

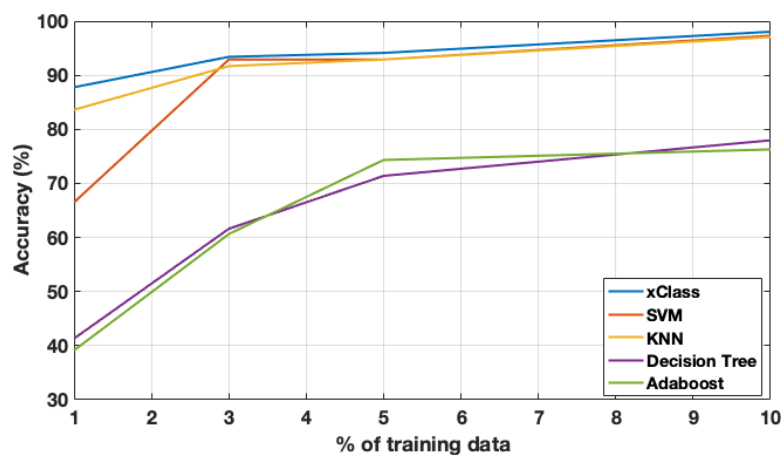


Fig. 12 Classification curve for different number of training samples for the Caltech-101 dataset.

proaches if just one class is used in the training phase). However, the proposed xClass approach could obtain better results in terms of accuracy even though it uses an extremely weak supervision (Fig. 14).

Fig. 13 illustrates the drop of confidence when new unseen classes are presented to the proposed classifier. Differently from traditional approaches which require the knowledge of all available classes beforehand, the proposed xClass uses its exploratory mechanism to autonomously discover this new class with basis just on the empirical data. Red bars on Fig. 14 shows the results obtained by state-of-the-art methods if just one class is presented during the training phase, as they are not able to detect new arrivals data patterns and

adapt their structure to this scenario, they wrongly classify the new arrived data samples as the known class. Different types of supervision (extremely weak, weak, full) is provided during experiments, in all cases the proposed method could provide better results in terms of classification performance than its competitors as illustrated by Fig. 15. It is possible to note through Fig. 14 that the results obtained for extremely weak supervision with xClass surpass its competitors in more than 25% in terms of classification performance, which indicates the efficiency of the proposed method.

As given by Fig. 15, xClass is able to improve its results if more training data and all classes are provided. For validation purposes, 20% of the data samples were used in all cases and labels for newly detected classes by xClass are attached during this phase. The AnYa fuzzy rule [6] for the newly identified class R_{new} can be written as follows:

$$R_{new} : \text{IF } (x \sim \begin{bmatrix} 104 \\ 41 \\ 66 \\ 10 \\ 23 \\ 635 \\ 73 \end{bmatrix}) \text{ THEN 'NewClass1'}$$

where x is the set of selected features given by the density-based feature selector. x can be written as follows:

$$x = \begin{bmatrix} \text{COMPACTNESS} \\ \text{CIRCULARITY} \\ \text{PR. AXIS ASPECT RATIO} \\ \text{MAX. LENGTH ASPECT RATIO} \\ \text{PR. AXIS RECTANGULARITY} \\ \text{SCALED VARIANCE MINOR} \\ \text{SKEWNESS ABOUT MAJOR} \end{bmatrix}$$

During the validation stage labels are attached to the newly identified rules. The final format for the first newly identified rule is given as follows:

$$R_2 : \text{IF } (x \sim \begin{bmatrix} 104 \\ 41 \\ 66 \\ 10 \\ 23 \\ 635 \\ 73 \end{bmatrix}) \text{ OR } (x \sim \begin{bmatrix} 90 \\ 34 \\ 55 \\ 6 \\ 17 \\ 224 \\ 65 \end{bmatrix}) \text{ OR...OR } (x \sim \begin{bmatrix} 113 \\ 53 \\ 62 \\ 11 \\ 24 \\ 688 \\ 72 \end{bmatrix}) \text{ THEN 'SAAB'}$$

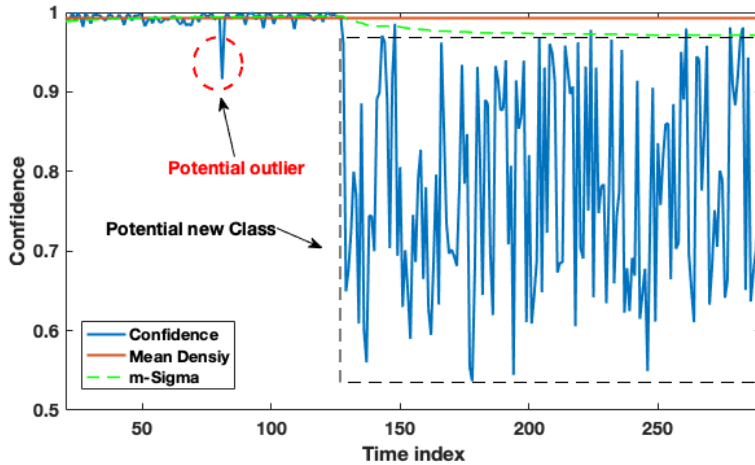


Fig. 13 Sudden drop of confidence due the presentation of new unknown classes – Cars dataset.

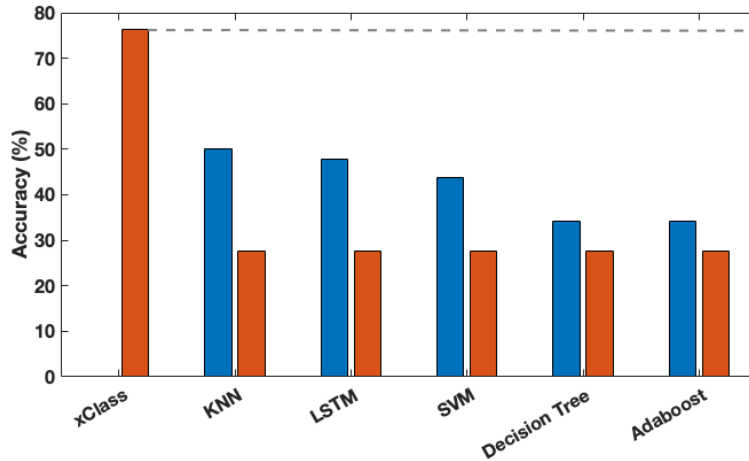


Fig. 14 Accuracy for extremely weak supervision classification for the Cars dataset.

4 Conclusion

In this paper, we break with the traditional approach to supervised classification. We offer a new fully autonomous extremely weakly supervised approach (xClass) which is able to learn from just a single class and a handful of labeled data samples. Then, as new classes, unknown to the human user the trained classifier appear at a later stage, the proposed xClass method is able to successfully discover this and learn from the data autonomously as demonstrated in the results section. Furthermore, the features (inputs to the classifier) are

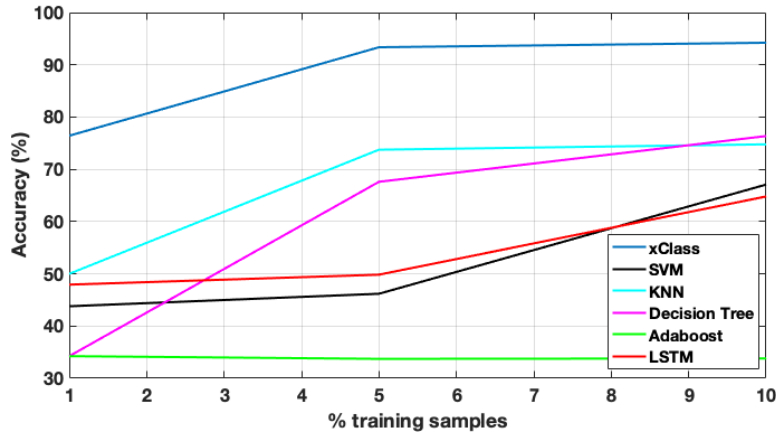


Fig. 15 Classification curve for different number of training samples for the Cars dataset.

automatically sub-selected by the algorithm based on the accumulated data density per feature per class. Results demonstrated that the proposed approach offers a high precision as compared with the top state-of-the-art algorithms.

The proposed xClass approach could surpass its competitors in terms of accuracy for all experiments using extremely weak supervision, as well as, full supervision. Moreover, the proposed algorithm produced highly transparent interpretable results, which are helpful for human experts analysis. No user- or problem- specific algorithmic parameter (except for feature selection which can be done by *ad hoc* decision) are required which is also an advantage provided by the proposed xClass classifier.

To validate our proposal we tested it on four challenging problems, including adversarial autonomous cars scenarios classification, imbalanced faces detection, and objects detection. **Not only we achieved higher accuracy (in one of the problems outperforming by 25% the other methods), but, more significantly, we only used the knowledge of just a single class beforehand and extremely weakly labeled data and we generated interpretable models with smaller number of features used. Furthermore, the proposed xClass method demonstrated the ability to learn from unknown without retraining, which is one of the biggest problems of deep learning based on neural networks. As illustrated the convolutional deep learning misclassified an unknown class with high confidence, on the other hand, the proposed approach was able to detect a sudden drop in the confidence and learn from this unknown data, then it was able to proactively create a new class for this new scenario. The proposed method is applicable to a wide range of problems, especially for problems with unknown dimension and for problems for which the concept changes over time.**

As a future work, we will investigate the occurrence of more than one unknown class at the same time. Furthermore, we will also explore highly dynamic problems such as video and other forms of data streams and address the time needed to learn online.

Competing Interests

The authors declare no competing interests.

References

1. Allen, G., Chan, T.: Artificial intelligence and national security. Belfer Center for Science and International Affairs Cambridge, MA (2017)
2. Angelov, P.: Autonomous learning systems: from data streams to knowledge in real-time. John Wiley & Sons (2012)
3. Angelov, P., Filev, D.P., Kasabov, N.: Evolving intelligent systems: methodology and applications. John Wiley & Sons (2010)
4. Angelov, P., Soares, E.: Towards explainable deep neural networks (xdnn). *Neural Networks* **130**, 185–194 (2020)
5. Angelov, P.P., Gu, X.: Toward anthropomorphic machine learning. *Computer* **51**(9), 18–27 (2018)
6. Angelov, P.P., Gu, X.: Empirical approach to Machine Learning. Springer (2019)
7. Angelov, P.P., Gu, X., Príncipe, J.C.: A generalized methodology for data analysis. *IEEE Transactions on Cybernetics* (99), 1–13 (2017)
8. Angelov, P.P., Zhou, X.: Evolving fuzzy-rule-based classifiers from data streams. *Ieee transactions on fuzzy systems* **16**(6), 1462–1475 (2008)
9. Bezerra, C.G., Costa, B.S.J., Guedes, L.A., Angelov, P.P.: A new evolving clustering algorithm for online data streams. In: 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 162–168. IEEE (2016)
10. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2722–2730 (2015)
11. Core, M.G., Lane, H.C., Van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.: Building explainable artificial intelligence systems. In: AAAI, pp. 1766–1773 (2006)
12. Faces, C.: Computational vision at caltech (1999)
13. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding* **106**(1), 59–70 (2007)
14. Gibney, E.: Deepmind algorithm beats people at classic video games. *Nature* **518**(7540), 465–466 (2015)
15. Hodgson, G.M.: The great crash of 2008 and the reform of economics. *Cambridge Journal of Economics* **33**(6), 1205–1221 (2009)
16. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Data preprocessing for supervised learning. *International Journal of Computer Science* **1**(2), 111–117 (2006)
17. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–958. IEEE (2009)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
19. Long, Y., Shao, L.: Learning to recognise unseen classes by a few similes. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 636–644 (2017)
20. Lughofer, E.D.: Flexfis: A robust incremental learning approach for evolving takagi-sugeno fuzzy models. *IEEE Transactions on fuzzy systems* **16**(6), 1393–1410 (2008)
21. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial tessellations: concepts and applications of Voronoi diagrams, vol. 501. John Wiley & Sons (2009)
22. Rezaei, M., Terauchi, M.: Vehicle detection based on multi-feature clues and dempster-shafer fusion theory. In: Pacific-Rim Symposium on Image and Video Technology, pp. 60–72. Springer (2013)
23. Robles, R.J., Kim, T.h.: Applications, systems and methods in smart home technology: A. *Int. Journal of Advanced Science And Technology* **15** (2010)
24. Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1), 73–79 (2011)

25. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5), 206 (2019)
26. Russ, J.C.: *The image processing handbook*. CRC press (2016)
27. Schmidhuber, J.: Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015)
28. Shah, S.H., Yaqoob, I.: A survey: Internet of things (IoT) technologies, applications and challenges. In: *2016 IEEE Smart Energy Grid Engineering (SEGE)*, pp. 381–385. IEEE (2016)
29. Škrjanc, I., Iglesias, J., Sanchis, A., Leite, D., Lughofer, E., Gomide, F.: Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A survey. *Information Sciences* **490**, 344–368 (2019)
30. Soares, E.A., Angelov, P.P., Costa, B., Castro, M., Nagesh Rao, S., Filev, D.: Explaining deep learning models through rule-based approximation and visualization. *IEEE Transactions on Fuzzy Systems* (2020)
31. Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O.: A ‘non-parametric’ version of the naive bayes classifier. *Knowledge-Based Systems* **24**(6), 775–784 (2011)
32. Tyugu, E.: Artificial intelligence in cyber defense. In: *2011 3rd International Conference on Cyber Conflict*, pp. 1–11. IEEE (2011)
33. Waldrop, M.M.: Autonomous vehicles: No drivers required. *Nature News* **518**(7537), 20 (2015)
34. Wang, F.Y., Zhang, J.J., Zheng, X., Wang, X., Yuan, Y., Dai, X., Zhang, J., Yang, L.: Where does alphago go: From church-turing thesis to alphago thesis and beyond. *IEEE/CAA Journal of Automatica Sinica* **3**(2), 113–120 (2016)
35. Xu, K., Wang, X., Wei, W., Song, H., Mao, B.: Toward software defined smart home. *IEEE Communications Magazine* **54**(5), 116–122 (2016)