To what extent should machine learning be supervised rather than unsupervised?

Guillaume Macneil

2020
April

Abstract

The contemporary debate between the use of supervised, semi-supervised and unsupervised machine learning has always been a very subjective and inconclusive topic of conversation. This dissertation aims to compare the capabilities of each machine learning solution to determine which is the most accessible and effective machine learning method. Based on analyses of the accuracy, accessibility and general capabilities of these machine learning solutions, it was determined that supervised machine learning is currently the best choice and will likely continue to be for the near future.

1 Introduction

To introduce you to the world of supervised and unsupervised machine learning, allow me to present to you a scenario. You are in charge of a pie assembly line in a prestigious pie factory, you have been presented with an interesting new proposition that provides two new opportunities: **1.** you could use a mix of human and machine, with the machines observing the human workers for the first month, but after that point the machines would manufacture the pies with 90% efficacy, **2.** you could venture into an unknown pie manufacturing paradigm – an assembly line with no human interference, one that relies purely on the machinery learning the way to create pies, using very non-human like methods to do so. Though there will be mistakes, with filling occasionally coating the floor and casings from time to time becoming dangerous projectiles, it will improve over time and does not require a salary to keep it satisfied in its servitude. Which would you pick?

Now, the above scenario was a (not so subtle) demonstration of the contemporary debate between the usage of supervised (**1.**) and unsupervised (**2.**) machine learning. There are certainly valid arguments in favour of both sides

Written with LaTeX.
With thanks to Ms. Daley and Mr. Williams.

of the debate. Deciding which side of the discussion is the most convincing through a (mostly) technological and (less so) economical lens is the aim of this dissertation. I will begin my dissertation by discussing the intricacies of supervised machine learning – its benefits, downsides, areas of excellence and areas in which it is not the best machine learning choice. I will then go on to discuss the alternatives under the same analytical light – unsupervised machine learning and semi-supervised machine learning (an approach that is not as easily demonstrated by my pie manufacturing analogy). I will then compare the three machine learning methods to determine which method is the most frequently applicable. To conclude my dissertation, I will summarise my findings and offer some potential paths that future developments in this field could follow and the time-frames that these changes are likely to follow.

For a better understanding of some of the more technical terms used in this dissertation, here are definitions for the key terms that will be used:

- **Machine learning** – The use of algorithms and statistical models by computer systems to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- **Supervised machine learning** – A way of writing algorithms that decides on an output based on a specific input, 'learning' how to decide this based on input-output decisions (labelled data) made by humans.
- **Unsupervised machine learning** – A way of writing algorithms that decides on an output based on a specific input, 'learning' how to make these decisions by finding previously unknown patterns without any labelled data.
- **Semi-supervised machine learning** – A compromise between supervised and unsupervised machine learning. A way of writing algorithms that uses a small amount of labelled data but also employs unsupervised learning techniques.

2 Main Findings

2.1 A Human-Led Approach - Supervised Machine Learning

Supervised learning is the most common branch of machine learning [Castle, 2017][Marr, 2017], involving the use of labelled data to train a model to follow the example of something that can already perform the task. Due to the technique's popularity, we find supervised machine learning approaches being applied to a diverse range of tasks, across many scientific disciplines, from predicting the social interactions between Rodentia [Hong et al., 2015] to identifying certain morphological features of galaxies [Kuminski et al., 2014] or even to code the main policy issues of newspapers and parliamentary questions [Burscher et al., 2015]. This approach to machine learning often has a very high success rate which is partially down to the human influence that is present in this machine learning method. These factors make supervised machine learning a very attractive approach for data scientists and researchers in general, but there are many instances in which supervised learning is not the optimal choice.

One of the most attractive attributes of supervised machine learning is its very high success rate or accuracy. In the calculations below, a measure called the 'f1 score' is used as it incorporates the precision and recall abilities of the machine learning algorithm in order to determine its accuracy [Sasaki, 2007]. 100% is the highest possible score, 0% is the lowest possible score. Table 1 shows the mean percentile f1 score of each of the supervised learning solutions that were used in the academic papers reviewed over the course of this dissertation. The average across all of the academic papers is 73.74%. This is a very high accuracy for any machine learning solution. In both 'Coding policy issues' and 'Predictions of rodent social interactions', these solutions were more accurate than the industry standard at that time. Krippendorff's alpha method for coding achieved 64.5% accuracy, where the supervised learning method achieved 68.4%. As stated in the 'Predictions of rodent social interactions' paper [Hong et al., 2015], there is 'a lack of automated, quantitative, and accurate assessment of social behaviours in mammalian animal models' and therefore their proposed solution is remarkably good with an f1 score of 79.1%. This high precision makes it well-suited to tasks that require a very low failure rate, but still have margin for error.

Table 1:

Use case:	f1 score /%:	Mean f1 score /%:	Total f1 score /%:
Coding: 10%-word use.	68.50	68.38	73.74
Coding: 100%-word use.	70.00		
Validation: 10%-word use.	67.50		
Validation: 100%-word use.	67.50		
Rodent int.: Attack	77.64	79.10	
Rodent int.: Investigation	81.67		
Rodent int.: Mounting	77.98		

This table shows the mean accuracy of a representative selection of supervised machine learning applications from two academic papers that were reviewed. The rows titled 'Coding' and 'Validation' show the f1 scores from the paper titled 'Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?' and the rows titled 'Rodent int.' (standing for rodent interaction) represent the f1 scores from the paper titled 'Automated measurement of mouse social behaviours using depth sensing, video tracking, and machine learning.'

Another attractive attribute of supervised machine learning is the method's relative simplicity. All of the academic papers that were reviewed followed the same basic method of training and application (and conceptually, all supervised machine learning algorithms do): **1.** determine certain features from a dataset that can be interacted with by the machine, **2.** hand label these features - showing what the desired result is, **3.** train the model on this labelled data over a certain number of epochs (repeats or cycles) in order to determine certain decision boundaries, **4.** use this model on other data. This is mentioned in (the names are shortened) 'Coding Policy Issues' [Burscher et al., 2015], 'Morphological Analysis of Galaxy Images' [Kuminski et al., 2014], 'Scoring Diverse Cellular Morphologies' [Jones et al., 2009] and 'Measurement of Mouse Social Behaviours' [Hong et al., 2015]. This method is not mentioned in 'Locating Damage in a Rotating Gear' [Kaul and Oza, 2005] but it is highly likely that it does. The source code used for the 'Morphological Analysis of Galaxy Images' (from a project named Wndchrm) is open-source and publicly available and uses 2Gb of RAM per core, a 2GHz CPU and 10Gb of hard drive space to run [L et al.], all of which is easily attainable on a cheap, commercially available personal computer. This means that almost anyone (within reason) can run an advanced supervised machine learning algorithm to solve a problem, whereas it was once a task only supercomputers could perform.

A notable downside of supervised machine learning is the data labelling phase of the algorithm's training process. This process requires a team of specialised people to hand-assign each piece of training data with a desired outcome. For example, in the case of 'Morphological Analysis of Galaxy Images' [Kuminski et al., 2014] a spherical or ovoid galaxy with discs and no sign of spiral arms would be annotated with either 'elliptical' or 'dwarf elliptical'. This is not necessarily a bad thing, however. Such definite data provides the algorithm with specific rules to follow and thus an absolute decision boundary from which to form certain decisions (like classification) and is primarily responsible for the high accuracy of the approach. However, this process is **very** intensive and greater accuracy requires more labelling. This is evident in 'Measurement of Mouse Social Behaviours' [Hong et al., 2015] which managed to achieve an f1 score of 79.10%, but only after training the model on 150,000 hand-labelled frames and validating the model on 350,000 hand-labelled frames. Therefore, for any large-scale supervised machine learning solution which requires a high (or even average) level of accuracy, such implementations would be limited by the time consuming and costly process of hand-labelling.

Another important consideration that could become a deciding factor between the usage of supervised, unsupervised or semi-supervised learning is supervised machine learning's relative inability to generalise across contexts due to its reliance on human-labelled data. This is most clearly evident in the paper 'Coding Policy Issues' [Burscher et al., 2015] as this paper was used to test whether supervised machine learning could generalise across contexts. They came to the conclusion that 'The ability of an supervised machine learning model to generalize across contexts, however, is limited and depends on the characteristics of available training data'. The paper proves this by testing algorithms that were trained with certain parameters on texts that have other parameters, for example, using an algorithm that was trained on Parliamentary Questions for coding newspapers. Table 2 is taken from this paper and shows how the algorithms performed in each situation in terms of f1 score. It is evident that, when compared to the baseline, all other implementations that generalise across contexts perform noticeably worse. The accuracy decreases in certain categories ($VK/NRC \rightarrow TEL$ or $2004 - 2011 \rightarrow 1995 - 2003$) though this may still be serviceable in certain applications, it is not ideal. This means that situations in which an algorithm is required to be particularly flexible in terms of the data it parses, supervised machine learning may not be applicable.

Table 2:

Baseline:		
$News \rightarrow News$	$PQs \rightarrow PQs$	
0.67	0.68	
Other Text:		
$News \rightarrow PQs$	$PQs \rightarrow News$	
0.50	0.49	
Other Newspaper:		
$VK/TEL \rightarrow NRC$	$NRC/TEL \rightarrow VK$	$VK/NRC \rightarrow TEL$
0.59	0.63	0.65
Other Time Frame:		
$1995 - 2003 \rightarrow 2004 - 2011$	$2004 - 2011 \rightarrow 1995 - 2003$	
0.59	0.63	

This is a table from 'Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?'. 'Baseline' illustrates algorithms being used on the type of texts that they were trained on. 'Other Text Sort', 'Other Newspaper' and 'Other Time Frame' shows algorithms being used on types of text they were not trained on. NB: VK = Volkskrant, NRC = NRC/Handelsblad, TEL = Telegraf (Dutch publications).

2.2 A Machine-Led Approach - Unsupervised Machine Learning

Unsupervised machine learning is an alternative approach to machine learning, involving a completely unlabelled dataset which is used to train the model, through the model's statistical analysis of the data, to separate each the data into groups called 'clusters'. Though unsupervised machine learning isn't as widely used as supervised machine learning [Castle, 2017][Marr, 2017], we can still see this method being used in the context of learning vowel categories through infant-directed speech [Vallabha et al., 2007] or the identification of the optimal control strategy for hybrid cars in terms of fuel efficiency (and other factors) [Finesso et al., 2016]. Though this approach has a noticeably decreased accuracy when compared to supervised machine learning, there is no need for any labelling of the input dataset leading to a vast efficiency increase. It is this balance of benefits and downsides that will be discussed.

The primary benefit of unsupervised machine learning is its (almost) complete independance from researchers, due to the lack of a labelled dataset as the input for the algorithm. This is acknowledged in the paper titled 'Unsupervised Training of Bayesian Networks for Data Clustering' [Pham and Ruz, 2009] in which it is stated that 'In many industrial applications of machine learning, it is difficult to obtain a large dataset with classified examples. This is generally to the fact that a human expert is needed to manually classify each example'. The economic issue which arises with the usage of supervised learning is remedied by the usage of unsupervised learning, as there is no need to label the dataset before training. This is evident in the case of 'Unsupervised Learning of Natural Languages' [Solan et al., 2005] in which the algorithm was successful in 'learning complex syntax' and 'generating grammatical novel sentences' with the only input being an unlabelled English language corpus. This leads to a vast increase in the efficiency of the solution, as it almost completely removes the user inefficiencies (the user still has to analyse the output at the end as it is not labelled, but this shall be discussed later). This therefore means that unsupervised machine learning could be preferable in situations in which a method of data analysis needs to be discovered quickly.

Another benefit of unsupervised machine learning is partially due to the unlabelled dataset input, it is that unsupervised machine learning can come to conclusions that a human could not, as it is not 'confined' to following the examples set by humans. This is most evident in the paper 'Unsupervised Learning of Vowel Categories from Infant-Directed Speech'[Vallabha et al., 2007] in which the researchers attempt to create an algorithm that detects vowel categories of English and Japanese based on infant-directed speech. This experiment used 20 English speakers and 10 Japanese speakers who pronounced nonce words to their infants, these were recorded and used as the dataset. In this case, the researchers do not exactly know how to learn the vowel categories from infant-directed speech and it is even stated that 'Japanese infants discriminate the English /**r**/ and /**I**/ that are confused by Japanese adults' [PK et al., 2006], so this field is evidently an area that is confusing for the researchers involved. However, this algorithm is uniquely qualified to analyse and parse this infant-directed speech as it doesn't need labelled data in the dataset. This is to say that unsupervised machine learning algorithms are well suited (and often the only machine learning option) for situations where the user doesn't completely understand how to manipulate the dataset.

There are however some significant issues with unsupervised machine learning. Arguably the most significant downside of unsupervised machine learning is its reduced accuracy when compared to supervised machine learning. This is stated in the paper titled 'Unsupervised Training of Bayesian Networks' [Pham and Ruz, 2009] from Cardiff University, in which it is stated that 'a clustering accuracy of 65 per cent might seem low. The best classification accuracy obtained for the same dataset in another study [Ruz and Estevez, 2005] using carefully selected and finely tuned multi-layer perceptron (MLP) neural networks was only 83 per cent'. This is a whole 18% difference between these two methods,which in practice would be a significant performance decrease. This is also supported by a paper ('Comparison Between Supervised and Unsupervised Classifications of Neuronal Cell Types: A Case Study')[Guerra et al., 2010] that directly compares multiple methods of both supervised (classification) and unsupervised (clustering) machine learning in the context of neuronal cell types. The average of the clustering algorithms (shown in Table 3) achieved an accuracy of 69% where the classification algorithms achieved an accuracy of 85% - a difference of 16%.This is a somewhat flawed paper as the unsupervised classification category only has 1 algorithm whereas supervised classification has 5 algorithms. This means that if the use case for a machine learning algorithm is highly accuracy-dependent (traffic light timing for example), unsupervised learning would not be applicable. Furthermore, the paper seems to have a somewhat flippant attitude towards the difference of 18% between their clustering accuracy and the neural network accuracy (in the usage of 'only 83 per cent') which may cast some doubts upon how valid the researchers' views are upon the difference in accuracy between the two methods.

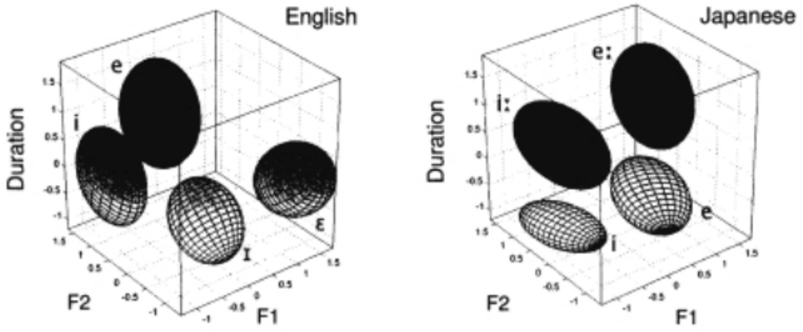
Table 3:

Classification:	Algorithm:	Accuracy /%:	Mean Accuracy /%:
Supervised	Naïve Bayes	82.05	84.88
	Decision Tree	85.32	
	K Nearest Neighbors	85.76	
	Multilayer Perceptron	86.24	
	Logistic Regression	86.02	
Unsupervised	Hierarchical Clustering	68.98	68.98

This is a table that collates all of the tables in 'Comparison Between Supervised and Unsupervised Classifications of Neuronal Cell Types: A Case Study' to determine a mean accuracy of supervised and unsupervised machine learning.

Another downside of unsupervised machine learning is the necessity for the user to analyse the output of the unsupervised machine learning algorithm due to the fact that the input dataset for these algorithms is completely unlabelled and therefore the output cannot be labelled as the cluster 'names' are not defined. This problem is well illustrated in the academic paper titled 'Unsupervised Learning of Vowel Categories from Infant-Directed Speech' [Vallabha et al., 2007], the output of this experiment was multiple vowel groupings (/i,e/ from /i:,e:/ in Japanese and /i,e/ from /I,E/ in English) from an unlabelled dataset of nonce words spoken by English and Japanese mothers. This paper illustrates the output clusters of this algorithm with graphs (visible in fig. 1). They are labelled with the vowel categories that they represent, which in this case would be relatively simple to determine simply by listening to the nonce words which are represented in the cluster, but could become an efficiency-decreasing factor depending on the number of clusters there are. It is pertinent to mention that this process of cluster analysis would take less time to perform than the labelling of a dataset, but it is still a factor to take into consideration. In practice, this process of cluster analysis could represent a significant hurdle in the development of the final solution.

Fig. 1

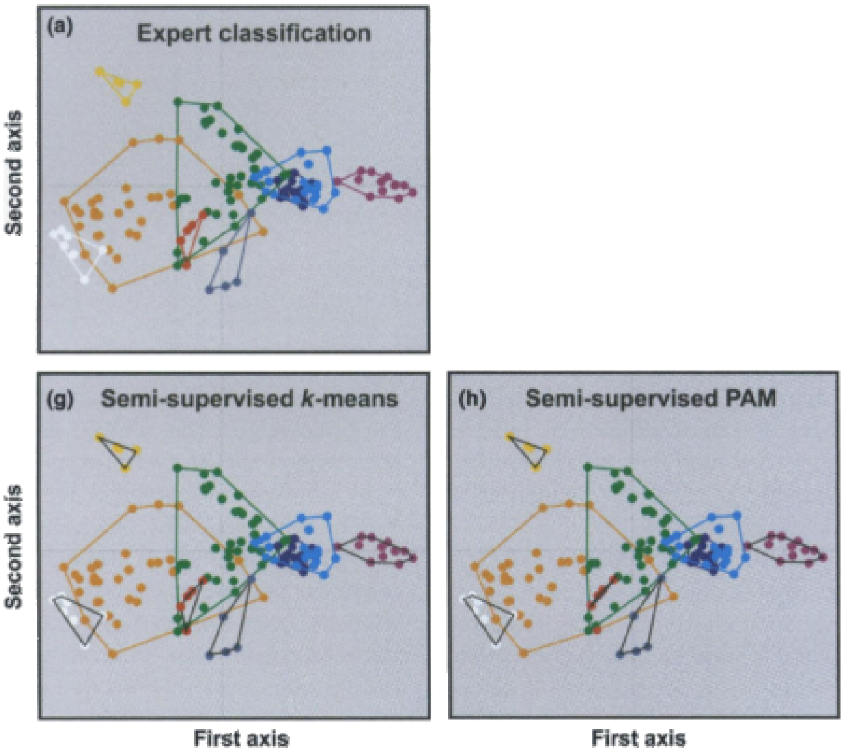


Graphs from 'Unsupervised Learning of Vowel Categories from Infant-Directed Speech' [Vallabha et al., 2007] that demonstrate the 'clusters' of data (in this case, they are vowel categories) that arise from unsupervised machine learning.

2.3 An Alternative Approach - Semi-supervised Machine Learning

With many of the papers that were analysed for this dissertation, it seems to be portrayed that there are only two approaches for machine learning - supervised and unsupervised - this is however a false dichotomy as there is the other promising option of semi-supervised machine learning. The major benefit of semi-supervised machine learning is due to the fact that it is composed of both a supervised and unsupervised algorithm and its usage of a small amount of labelled and a large amount of unlabelled data. These conditions allow the semi-supervised algorithm to maintain a relatively high accuracy while also being able to generalise over contexts remarkably well. This is clearly displayed in the paper titled 'Semi-supervised classification of vegetation' [Tichý et al., 2014], in this experiment the algorithm (a modified k-means algorithm used along side a PAM algorithm) was tasked to classify certain 'community ecology and vegetation' and identify new types of vegetation. A traditional supervised approach could not solely be relied upon as the algorithm is required to identify 'new group sites that do not fit well into the a priori groups' which requires generalisation. Unsupervised learning could not be used either as each new classification leads to 'partitions that are partially inconsistent with previous classifications' making classification inaccurate. As can be seen in Fig. 2 the semi-supervised algorithm 'identified almost perfectly the remaining five expert based groups that were not defined a priori', the graphs produced by the algorithm are virtually indistinguishable from the (human) expert classification. This means that semi-supervised algorithms lend themselves well to situations in which the ability to generalise across contexts is needed while accuracy is still valuable.

Fig. 2 [Tichý et al., 2014]



An example of semi-supervised machine learning's high accuracy and ability to generalise. These graphs were taken from 'Semi-supervised classification of vegetation: preserving the good old units and searching for new ones' and represents the semi-supervised algorithm's ability to classify '144 forest vegetation plots'.

Another significant benefit of semi-supervised machine learning is that only a small quantity of labelled data is needed to achieve a high accuracy and an addition of excess data (over-sampling) can result in diminishing returns. This was shown in a paper titled 'Using semi-supervised classifiers for credit scoring.'[Kennedy et al., 2013] in which a semi-supervised algorithm was used to determine the probability of default for low-default portfolios (LDPs) for which very little history exists. In these circumstances, it was determined that the type of semi-supervised algorithm that they used (one-class classification algorithms) should only be relied upon 'in the near or complete absence of defaulters', so this implementation was not ideal. It was however determined that the process of oversampling (introducing a bias to select more samples from one class than the other)[Jesús et al., 2015] produced 'no overall improvement' when compared to the best 'two-class classification algorithms' which were the solution prior to the experiment using semi-supervised machine learning. This means that semi-supervised learning could provide a good alternative to unsupervised learning for situations in which the accessible data is mainly unlabelled but a small amount of labelled data could be acquired, as it would likely lead to an increase in the accuracy of the solution.

However, a notable downside to semi-supervised machine learning is that this method of machine learning still relies on the human who is labelling the small portion of labelled data to know how to process the data themselves (due to the very nature of the process of hand labelling). In the academic paper titled 'Shock Waves of Political Risk on the Stock Market: The Case of Korean Companies in the U.S.' [Pak et al., 2015] a semi-supervised algorithm was used to determine the effect that 'North Korean risk' had on the NYSE and the KOPSI (the South Korean equivalent to the New York Stock Exchange). In this case, 3 human coders were employed to classify 2000 articles with either negative, neutral or positive sentiments and this was used as part of the training set for the algorithm. Between the human classifiers, 55.3% of assigned sentiments were shared by all of the coders, and 44.3% of the sentiments were shared by two of the coders. This level of agreement between the coders is relatively high, but in applications in which hand-labelling is more complex and the human coders are less experienced (you may recall the paper 'Combining Human and Machine Learning for Morphological Analysis of Galaxy Images.' [Kuminski et al., 2014] in which each coder had to answer 10 questions and agreements were occasionally as low as 50%), this may be a large contributing factor to a low accuracy. This means that semi-supervised learning is restricted to applications in which supervised learning can be used and therefore may not always be a desirable approach when supervised alternatives are considered.

Another limitation with semi-supervised machine learning, particularly in complex applications, is that this approach to machine learning requires a certain amount of assumptions to be made about the unlabelled dataset based on the labels provided by the labelled dataset. It is important to not confuse 'assumption' and 'inference' in this context, as it is desirable for the semi-supervised model to make inferences about the unlabelled dataset based on the labelled dataset so that parameters determined from the labelled dataset can be projected onto the unlabelled dataset. However, assumptions are not desirable as, in this case, they are universally applied and can therefore lead to an incorrect analysis of the data. The utopian semi-supervised algorithm would be one that makes no assumptions about the relationship between the labels and the unlabelled dataset as it could be damaging. This concept is called a 'no-prior-knowledge' dataset [Lu, 2009]. Unfortunately, this ideal is currently unachievable since assumptions are inherent to the nature of semi-supervised algorithms as labelled data has to be inputted for the model to interpret the unlabelled data. The people labelling the data will inherently make assumptions on the relationship between the labels and the unlabelled dataset while generating the labelled sample. The most important assumptions (as listed in 'Fundamental Limitations of Semi-Supervised Learning' [Lu, 2009]) are listed below:

- **The cluster assumption.** – "If points are in the same cluster, they are likely to be of the same class."

  This seems like a very reasonable assumption - if a number of data points are clearly grouped together then it is highly likely that there is some relation between them. The problem arises with the definition of a cluster. Though it is easy for a human to see a high density grouping of points and define it as a cluster, it is a complex task for a machine learning algorithm to perform.

- **Low density boundary assumption.** – "The decision boundary should lie in a low density region."

  Upon first glance, this also seems like a reasonable assumption. However, this statement also falls down due to the complications in defining 'low density'. Though it is easy for a human to compare the densities of points in a distribution, certain questions arise in the case of attempting to do this with an algorithm - "Should the low density boundary be the lowest density boundary possible or should it have some density ϵ away from the lowest density boundary? What should this tolerance, ϵ, be?" [Lu, 2009]

These assumptions can lead to flaws in the model and, in the worst cases it is possible that the model can 'end up doing much worse than simply ignoring the unlabelled data and performing supervised learning' [Lu, 2009] making semi-supervised learning potentially undesirable in situations in which unlabelled data shares a difficult-to-discern trend or is simply low quality.

3 Evaluation

Supervised machine learning is a very popular choice [Castle, 2017][Marr, 2017] across many different data analysis disciplines and applications, but how deserved is this status? Due to the input to the algorithm being solely labelled data, the algorithm can achieve very high accuracies (see table 1). Furthermore, this type of labelled input requires very little pre-processing for it to be used by the algorithm. This leads to these algorithms being accessible and being comparatively less computationally intensive than their alternatives. However, it could be argued that this decrease in intesiveness is made up for in inefficiency by the necessity to hand-label data. Though in theory this is true, recently there has been an emergence in the creation of open-source, massive labelled datasets created by large corporations, which span many applications. There are datasets for hand-written digits (MNIST [LeCun et al.]), datasets for hand-drawn images (Quickdraw [Google]) and even datasets for deadly diseases like breast cancer (Breast Cancer Wisconsin Diagnostic Dataset [Wolberg], henceforth called BCWD). It is also pertinent to note that some datasets (like BCWD and MNIST) are painstakingly hand-labelled by volunteers but some others are cleverly labelled in other ways. Most notable of these datasets is the Google

Quickdraw dataset that labels its images by asking the user to play a game in which they draw 6 images based on prompts they are given. This has allowed that dataset to amass millions of labelled images since over 15 million people have played the game. However, supervised learning algorithms still cannot generalise well [Burscher et al., 2015], meaning that if you were to input an image of a cuckoo into a supervised learning model that was trained to classify pigeons,dogs and cats, the output would likely classify it as a bird, but not a cuckoo. This is due to the model following human-labelled data very closely.

Unsupervised machine learning takes a very different approach to data analysis to supervised machine learning and has many benefits too. Unsupervised machine learning only takes in unlabelled data as its input (for both training and testing) and therefore does not require any human-labelling which can be incredibly time consuming and project resource intensive (this is evident in the paper titled 'Automated measurement of mouse social behaviours using depth sensing, video tracking, and machine learning' [Hong et al., 2015] in which a total of 400,000 frames were hand labelled). So unsupervised learning could be far more time efficient and economically efficient. This benefit is becoming less and less valuable now that more open-source datasets are becoming widely available (as mentioned above). The other main benefit not offered by supervised machine learning is unsupervised learning's capacity to process data and come to conclusions which humans cannot intuitively draw, which is well documented by the academic paper that used unsupervised machine learning to learn certain vowel categories of English and Japanese through infant-directed speech [Vallabha et al., 2007]. There are significant shortcomings to unsupervised machine learning though, the most notable being its decreased accuracy when compared to unsupervised machine learning, a bi-product of the lack of a concrete decision boundary which is provided by human-labelled data. This is well demonstrated in the paper titled 'Unsupervised Training of Bayesian Networks for Data Clustering' in which a difference in accuracy of 18% was noted between the unsupervised model (65%) and a 'finely tuned multi-layer perceptron (MLP) neural network' which achieved (83%). Another problem with unsupervised machine learning is the necessity for a researcher to analyse the clusters produced by the model after it is applied to the data, this introduces a time consuming process which is not present in supervised or (certain) semi-supervised approaches.

Semi-supervised machine learning is an interesting compromise between supervised and unsupervised machine learning and provides a combination of the benefits and the downsides of the other two alternatives. Good implementations of semi-supervised learning are able to maintain a relatively high level of accuracy whilst being able to generalise over contexts. This gives semi-supervised learning a unique position in the comparison between supervised and unsupervised learning as it has a combination of positive attributes that both of its alternatives lack. This ability of semi-supervised learning is best illustrated in the paper titled ‘Semi-supervised classification of vegetation: preserving the good old units and searching for new ones' [Tichý et al., 2014], in which (as illustrated in Fig. 2) the algorithm was able to almost perfectly match a vegetation analysis performed by a human expert. Furthermore, the amount of labelled data necessary to be inputted into the model for an increased accuracy over unsupervised machine learning has been proven to be relatively small. A greater amount of labelled data used as an input has also been proven (in the paper titled 'Using semi-supervised classifiers for credit scoring' [Kennedy et al., 2013]) to yield diminishing returns in terms of accuracy increases. However, semi-supervised learning remains hampered by its reliance on a human being able to process the labelled data and therefore semi-supervised learning is of limited use in situations in which there has been no human precedent of solving the problem to which the model is applied. Semi-supervised models are also hampered by the necessity to make assumptions about the unlabelled dataset (like the 'cluster assumption' and the 'low density boundary assumption' [Lu, 2009]) which are not always appropriate and can be damaging to the accuracy of the model based on the quality of the unlabelled data.

4 Discussion

Supervised learning is an incredibly useful machine learning method for data scientists as it can achieve very high accuracies whilst also having a comparatively low processing power requirement. Supervised machine learning does struggle to generalise across contexts and does have a **currently** unavoidable dependance on humans for hand labelling. Unsupervised learning lies on the other end of the machine learning spectrum, with the main strength of this method being its independance from human beings. However this strength comes with the cost of accuracy and complexity. Semi-supervised machine learning is a tempting compromise for machine learning researchers as it manages to achieve a relatively high accuracy while also doing impressively well with generalisation task. However, it is held back by the neccesity for potentially dangerous assumptions to be made by the algorithm for it to function correctly.

Based on this, I believe that currently and for the next 5 to 10 years, supervised machine learning will remain the most widely used and widely applicable machine learning solution for both industry and general public usage. This is primarily due to the abundance of highly detailed datasets in most fields of usage, leading to a high degree of accuracy and performance while maintaining a relatively accessible level of complexity and intensity. It is pertinent to note that this does not rule out the usage of unsupervised or semi-supervised machine learning currently or in the near future. Perhaps in the distant future, unsupervised machine learning will usurp the other two alternatives due to its independance from humans. That said, for this to happen there would have to be sizable improvements in the field of unsupervised machine learning. At the point in which a unsupervised learning overtakes supervised and semi-supervised methods, unsupervised learning will theoretically be able to label datasets. At this point, it would be an unwarranted generalisation to think that datasets have to be hand-labelled, and at this point the boundaries of supervision between the methods begin to fall apart. It raises the question that if an unsupervised algorithm trains a supervised model, how is it classified?

5 Conclusion

To conclude, I believe that supervised machine learning is currently the most dominant type of machine learning and will continue to be for the next five to ten years, with the evolution of machine learning technologies following the timeline described above. That being said, one of the quotes that made the greatest impact on me through my analysis of academic papers discussing the fields of supervised, unsupervised and semi-supervised machine learning is this: 'The comparative assessment of classification methods can be a subjective exercise. It is influenced, among other factors, by the expertise of the user with each of the methods and the effort invested in refining and optimising each method'[Kennedy et al., 2013]. Therefore, I do not think that the selection of supervised machine learning as the currently dominant method of machine learning invalidates the current research and development in the other machine learning fields, I would even go so far as to sat that such a decision to ignore these technologies could be a grave mistake.

References

Bjorn Burscher, Rens Vliegenthart, and Claes H. De Vreese. Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The Annals of the American Academy of Political and Social Science*, 659:122–131, 2015.

Nikki Castle. Supervised vs. unsupervised machine learning, July 2017. URL https://blogs.oracle.com/datascience/supervised-vs-unsupervised-machine-learning.

Roberto Finesso, Ezio Spessa, and Mattia Venditti. An unsupervised machine-learning technique for the definition of a rule-based control strategy in a complex hev. *SAE International Journal of Alternative Powertrains*, 5:308–327, 2016.

Google. Google quick, draw! URL https://quickdraw.withgoogle.com/data.

Luis Guerra, Laura M. McGarry, Victor Robles, Concha Bielza, Pedro Larranaga, and Rafael Yuste. *Comparison Between Supervised and Unsupervised Classifications of Neuronal Cell Types: A Case Study.* Wiley, 2010.

Weizhe Hong, Ann Kennedy, Xavier P. Burgos-Artizzu, Moriel Zelikowsky, Santiago G. Navonne, Pietro Perona, and David J. Anderson. Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 112:E5321–E5360, 2015.

Julio Hernandez Jesús, Ariel Carrasco-OchoaJosé, and Francisco Martínez-Trinidad. Shock waves of political risk on the stock market: The case of korean companies in the u.s., 2015.

Thouis R. Jones, Anne E. Carpenter, Michael R. Lamprecht, Jason Moffat, Serena J. Silver, Jennifer K. Grenier, Adam B. Castoreno, Ulrike S. Eggert, David E. Root, Polina Golland, David M. Sabatini, and Edward M. Scolnick. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 106:1826–1831, 2009.

Upender K. Kaul and Nikunj C. Oza. Machine learning for detecting and locating damage in a rotating gear. *SAE Transactions*, 114:1198–1202, 2005.

K Kennedy, B Mac Namee, and SJ Delany. Using semi-supervised classifiers for credit scoring. *The Journal of the Operational Research Society*, 64:513–529, 2013.

Evan Kuminski, Joe George, John Wallin, and Lior Shamir. Combining human and machine learning for morphological analysis of galaxy images. *Publications of the Astronomical Society of the Pacific*, 126:959–967, 2014.

Shamir L, Orlov N, Eckley DM, Macura T, and Johnston J. Wnd-chrm github page. URL https://github.com/wnd-charm/wnd-charm.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. URL http://yann.lecun.com/exdb/mnist/.

Tyler (Tian) Lu. Fundamental limitations of semi-supervised learning. Master's thesis, University of Waterloo, 2009.

Bernard Marr. Supervised v unsupervised machine learning – what's the difference?, March 2017. URL https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#44e9016e485d.

Yunjung Pak, Young-Jin Kim, Min Song, and Yong-Hak Kim. Shock waves of political risk on the stock market: The case of korean companies in the u.s. *Development and Society*, 44:143–165, 2015.

Duc Truong Pham and Gonzalo A. Ruz. Unsupervised training of bayesian networks for data clustering. *Proceedings: Mathematical, Physical and Engineering Sciences*, 465:2927–2948, 2009.

Kuhl PK, Stevens E, Hayashi A, Deguchi T, Kiritani S, and Iverson P. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Institute for Learning and Brain Sciences and Department of Speech and Hearing Sciences*, 2006.

Gonzalo A. Ruz and Pablo Estevez. Image segmentation using fuzzy min-max neural networks for wood defect detection. *I*PROMS Virtual Conference*, 1: 6, 2005.

Yutaka Sasaki. The truth of the f-measure. Master's thesis, University of Manchester, 2007.

Zach Solan, David Horn, Eytan Ruppin, Shimon Edelman, and James L. McClelland. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, 102:11629–11634, 2005.

Lubomír Tichý, Milan Chytrý, and Zoltán Botta-Dukát. Semi-supervised classification of vegetation: preserving the good old units and searching for new ones. *Journal of Vegetation Science*, 25:1504–1512, 2014.

Gautam K. Vallabha, James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, 104:13273–13278, 2007.

Dr. William H. Wolberg. Breast cancer wisconsin (original) data set. URL https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).