Une approche visuelle pour l'exploration des résultats d'un grand nombre de classifieurs

Hélène Walle*, Pascal Makris*, Yassine Mofid**
Nadia Aguillon-Hernandez**, Claire Wardak**, Gilles Venturini*

*LIFAT, EA 6300, Université de Tours,
nom.prenom@univ-tours.fr, https://lifat.univ-tours.fr/
** UMR 1253, iBrain, Université de Tours, INSERM
nom.prenom@univ-tours.fr, https://ibrain.univ-tours.fr/

Résumé. Nous proposons une méthode visuelle pour analyser un grand nombre de résultats d'apprentissage supervisé, en l'illustrant sur une application réelle (détection des troubles autistiques à partir de données de eye tracking) dans laquelle nous avons pu obtenir plus de 30000 résultats de classification. Chaque triplet représentation-classifieur-paramétrage est positionné en 2D avec un algorithme de réduction de dimensions (MDS ou t-SNE) sur la base d'une distance utilisant la probabilité des classes prédites pour chaque donnée. Avec différents paramétrages de cette visualistion, nous montrons que l'utilisateur peut observer et évaluer visuellement l'efficacité des représentations, des classifieurs et de leur paramétrage.

1 Introduction

Lors de l'application d'un algorithme d'apprentissage supervisé à des données, les utilisateurs sont confrontés à des choix multiples à effectuer. Pour classifier des objets, il faut choisir des espaces de représentation traitables par un algorithme d'apprentissage (tables avec des descripteurs par exemple), et efficaces vis à vis de la tâche de classification (séparation des classes). Ensuite, il faut choisir un classifieur et son paramétrage, avec le même souci d'efficacité. Le choix final d'un triplet représentation-classifieur-paramétrage peut être simple lorsque peu d'options sont disponibles. Mais il peut arriver aussi que des milliers de variantes soient finalement générées en considérant les choix à faire pour l'espace de représentation, le classifieur et son paramétrage "optimal". Généralement, le choix du triplet représentation-classifieur-paramétrage le plus efficace se fait en considérant une mesure de qualité de la classification ("accuracy", AUC, autres mesures issues de la matrice de confusion), et en choisissant le triplet qui donne la meilleure valeur de l'indice. Cependant cette approche a une valeur explicative faible, et réduit beaucoup les informations que l'utilisateur pourrait obtenir : quelles représentations sont les meilleures (ou inefficaces) ? Certaines sont-elles équivalentes ou au contraire très différentes des autres dans leur manière de structurer l'espace de représentation des exemples ? Les mêmes questions peuvent se poser concernant les classifieurs et leur paramétrage.

Pour répondre à ces questions, nous adoptons une approche centrée utilisateur afin de favoriser au maximum l'exploration libre et les interactions. Nous étudions les approches visuelles

et interactives permettant de représenter les résultats de l'apprentissage supervisé. Les travaux antérieurs peuvent être classés de la manière suivante : le premier type d'approche consiste à se focaliser sur un type de classifieur en particulier, ce qui a pour avantage d'aller en profondeur dans la compréhension du fonctionnement du classifieur, comme dans Ghosh et al. (2005) (algorithmes de type KNN), Neto et Paulovich (2021) (pour RF), ou encore Chae et al. (2017) Rauber et al. (2017) Zintgraf et al. (2017) pour les réseaux de neurones.

Le deuxième type d'approches est générique car les méthodes peuvent considérer n'importe quel classifieur comme Schulz et al. (2015) Alsallakh et al. (2014) Luque et al. (2022), ou même un ensemble de classifieurs Meng et al. (2022). Elles donnent moins de détails sur le modèle appris, mais elles peuvent comparer plusieurs classifieurs entre eux. On trouve dans ces méthodes celles qui sont les plus proches de nos travaux : Alaiz-Rodríguez et al. (2008) proposent d'analyser de multiples classifieurs agissant sur différents jeux de données. Les performances de chaque classifieur ("accuracy", AUC, etc) servent de descripteurs pour construire une projection 2D des classifieurs (chaque classifieur est un point dans la représentation). Les visualisations regroupent donc ensemble une dizaine de classifieurs mais elles s'arrêtent au niveau jeu de données sans analyser les résultats en considérant les exemples des jeux de données. Heyen et al. (2020) ont développé le système ClaVis et proposent plusieurs visualisations des résultats d'une centaine de classifieurs. En particulier, ils proposent un nuage de points où chaque point est un classifieur et où une réduction de dimension permet de positionner en 2D les points. La mesure de distance entre deux classifieurs s'appuie sur les classes prédites et les probabilités des classes. Cet article cependant est limité dans le sens où il n'étudie pas en détail les propriétés de cette visualisation (appliquée sur la base des Iris, sans analyse particulière) et se concentre plutôt sur d'autres approches.

Dans notre proposition, nous étendons cette dernière visualisation d'un point de vue qualitatif et quantitatif (section 2), et dans des conditions expérimentales réelles (section 3) : nous allons considérer non seulement le choix des classifieurs et de leur paramétrage mais aussi celui de la représentation des données, avec des milliers de résultats de classification, dans le domaine de la détection automatique des Troubles du Spectre de l'Autisme (TSA).

2 Méthode proposée

L'approche que nous proposons consiste à construire une visualisation 2D dans laquelle chaque point représente un résultat de classification d'un triplet représentation-classifieur-paramétrage. Les variables visuelles associées à chaque point (forme, couleur, taille) dépendront de l'analyse à effectuer. Le positionnement 2D des points doit refléter les similarités qui existent entre les triplets, afin de mettre à jour des groupes de triplets au comportement classificatoire similaire, ou au contraire des points isolés classifiant différemment les données. Pour calculer ce positionnement, nous utilisons une matrice de distances entre triplets, cette distance permettant de mesurer les différences entre deux résultats de classification. Ensuite nous appliquons un algorithme de réduction de dimension de type "Multi-dimensional scaling" (MDS) ou t-SNE sur cette matrice pour obtenir le positionnement 2D des points.

Pour définir une distance entre résultats de classification, notre méthode utilise une matrice M de résultats de classification : les colonnes de la matrice sont les n objets à classer. Les lignes de la matrice sont les k classifieurs testés (k triplets représentation-classifieur-paramétrage). Une valeur dans cette matrice représente un vecteur de probabilités des classes. Pour des rai-

sons de clarté, et sans limiter la portée de notre approche, nous considérons un problème de classification binaire avec deux classes 0 et 1. On représente dans la matrice la probabilité de la classe 1. De plus, les n colonnes de la matrices peuvent être compléter par d'autres indicateurs permettant d'évaluer le résultat obtenu par un triplet : des mesures de qualité de la classification ("accuracy", "AUC", etc) ou encore les valeurs de la matrice de confusion. Ensuite, une distance euclidienne permet de comparer deux triplets.

3 Résultats

3.1 Contexte

Dans le cadre de notre projet SIRCUS (ANR-21-CE17-0045), notre objectif est de mettre au point un dispositif nomade et automatique permettant de détecter des TSA chez des enfants, à partir de données de eye-tracking. Le principe consiste à présenter des stimuli visuels (images ou vidéos) aux enfants afin de mesurer les réponses oculométriques (l'endroit où se pose le regard) caractérisant l'exploration visuelle des stimuli, et pupillométriques (diamètres des pupilles) caractérisant l'éveil physiologique en réponse aux stimuli. L'hypothèse sous-jacente pour la détection automatique des TSA est que les mouvements de l'oeil (coordonnées X Y) ou l'évolution du diamètre des pupilles ne sont pas les mêmes pour les enfants ayant des TSA par rapport aux enfants neurotypiques. Ces différences entre les deux groupes étudiés ont été distinguées par plusieurs articles comme celui de Guimard-Brunault et al. (2013) pour l'exploration oculométrique, ou celui de Aguillon-Hernandez et al. (2020) pour les variations pupillométriques. L'originalité de notre travail du point de vue de la détection des TSA est le fait que nous utilisons de nombreuses représentations différentes des données.

Dans notre protocole, 12 vidéos de 4 secondes chacune sont présentées à 170 enfants (79 TSA et 91 "Contrôle"). Pour chaque vidéo, nous mesurons avec une fréquence de 60Hz le point observé sur l'écran (coordonnées X Y du regard) ainsi que les diamètres des pupilles. Les 12 vidéos représentent des visages d'hommes et de femmes avec différentes expressions faciales (joie, tristesse, neutralité). Avec 170 enfants, nous obtenons théoriquement $170 \times 12 = 2040$ fichiers de données brutes, mais nous en avons en réalité 1833. Les enfants sont libres de leur mouvement et peuvent ne pas regarder une vidéo dans son ensemble, ou même ne pas la regarder du tout. Dans ce dernier cas, les classifieurs ne pourront pas décider, et la probabilité de la classe TSA produite en sortie est fixée à 0.5. Cela permettra de ne pas avoir de valeurs manquantes lors du calcul de distances entre résultats de classification.

Ces fichiers bruts contiennent principalement cinq colonnes : le temps, les valeurs X et Y de la position du regard, les deux diamètres des pupilles. Des pré-traitements ont lieu pour enlever les erreurs de mesure ou encore pour n'obtenir qu'une seule valeur pour le diamètre des pupilles. Ensuite de nouveaux descripteurs ayant déjà prouvé leur efficacité dans des études antérieures, sont ajoutés : dX et dY représentent le mouvement relatif des yeux d'un instant à l'autre, Théta représente en radians la direction prise par le regard à chaque instant.

Nous définissons ensuite quatre grandes manières de transformer ces données brutes en des représentations de type attributs-valeurs traitables par un algorithme d'apprentissage supervisé. Pour une vidéo donnée, nous pouvons construire :

— trois représentations oculométriques. A partir de X-Y ou dX-dY, nous calculons des histogrammes 2D ("heat maps"), avec différentes tailles de grille (5×4, 10×8, etc).

Chaque case de la grille compte combien de fois le regard (ou sa variation) est venu à cet endroit. Pour Théta, nous construisons des histogrammes 1D avec 8, 12, 16 ou 20 secteurs possibles,

— une représentation pupillométrique, avec deux variantes. Un histogramme 1D est calculé, soit directement avec le diamètre brut en mm, soit en normalisant ce diamètre avec les valeurs minimum et maximum du diamètre pupillaire de chaque enfant. Le nombre de boites ("bin") de l'histogramme peut être 10, 20 ou 30.

De plus, nous pouvons calculer ces histogrammes en effectuant des comptages, ou utiliser des fréquences, ce qui double le nombre de représentations possibles. Toutes ces représentations sont à multiplier par 12, car nous avons 12 vidéos présentées par enfant. Au total, nous obtenons avec ce processus 384 tables au format attributs-valeurs.

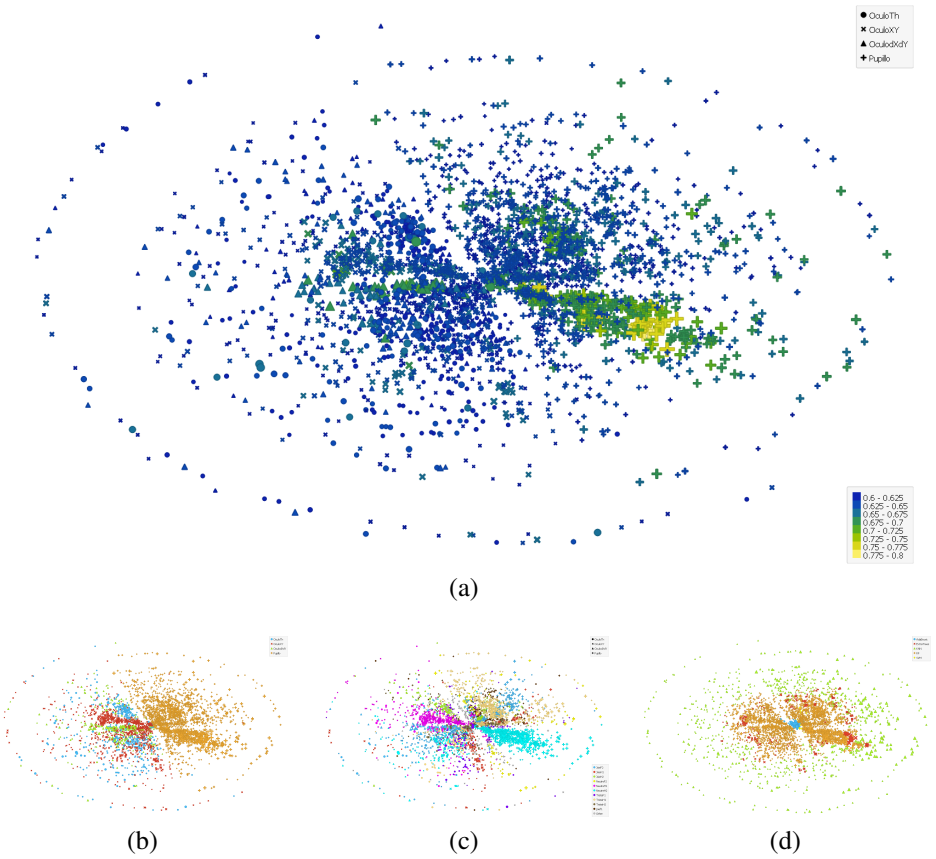
La dernière étape consiste à appliquer cinq types de classifieurs sur ces tables : RF, ExtraTrees, SVM, KNN, AdaBoost. Concernant les paramètres, nous testons plusieurs variantes pour chaque classifieur (voir section 3.3, avec 85 couples classifieur-paramétrage). En appliquant ces 85 classifieurs sur les 384 tables, nous obtenons 32640 résultats de classification, chacun décrit avec $n = 170$ valeurs de probabilités (de la classe TSA). Le problème posé dans l'introduction est donc concrètement le suivant : comment analyser et obtenir des informations pertinentes sur ces résultats ?

3.2 Vue d'ensemble

Nous avons trié les 32640 résultats par "accuracy" décroissante, les valeurs allant de 80% à 33%, et nous avons gardé 4000 résultats pour nos analyses (les résultat ayant une "accuracy" supérieure à 61%). Pour obtenir les visualisations, nous avons utilisé le logiciel Orange (orangedatamining.com).

Nous utilisons un algorithme de MDS pour obtenir une vue d'ensemble des données (voir figure 1). Concernant l'étude des représentations, on observe deux grands groupes de points dans cette visualisation : à gauche les représentations oculométriques et à droite les représentations pupillométriques (voir figure 1(b) où le type de représentation est codé comme une couleur). C'est dans cette partie "pupillométrique" du graphique que l'on trouve les meilleures performances de classification, un résultat attendu par les experts qui mettent en avant ces données sur la pupille (approche peu étudiée dans la littérature sur les TSA). La séparation des deux groupes vient des performances qui sont meilleures avec les données pupillométriques, mais aussi du fait qu'il y a plus de valeurs manquantes pour les données pupillométriques, car elles sont plus dures à acquérir pour certains enfants. Dans ce cas les classifieurs émettent une probabilité de 0.5, ce qui caractérise leur prise de décision sur certains enfants, et ce qui se traduit visuellement par un grand groupe à part.

Pour les experts le choix des vidéos est crucial, puisque l'on cherche celles qui sont les plus discriminantes. Dans la visualisation, les groupes de résultats se structurent beaucoup par vidéos (voir figure 1(c) où la vidéo est codée comme une couleur). Certaines sont très présentes (comme NeutreH2 qui inclut aussi les meilleurs résultats) et d'autres sont quasiment absentes des 4000 meilleurs résultats (comme NeutreF1). Cela indique aux experts que certaines vidéos sont plus intéressantes que d'autres, et suggère de ce fait une analyse particulière : pour construire ces vidéos, des acteurs ont été sollicités, hommes et femmes, et il se trouve que des éléments du visage ou de l'expression viennent influencer la qualité discriminante de la vidéo.



(a)

(b) (c) (d)

FIG. 1 – *Visualisation d'ensemble des résultats de 4000 triplets représentation-classifieur-paramétrage. Le positionnement des points est déterminé par un algorithme de MDS. La couleur et la taille représentent l'"accuracy", la forme code le type de représentation. Le nuage de points à gauche regroupe les représentations oculométriques, et celui de droite les représentations pupillométriques, comme cela est confirmé en (b) où la couleur code la représentation. En (c), nous indiquons quel stimulus est concerné parmi les 12 présentés. En (d), nous montrons comment se répartit le type de classifieur.*

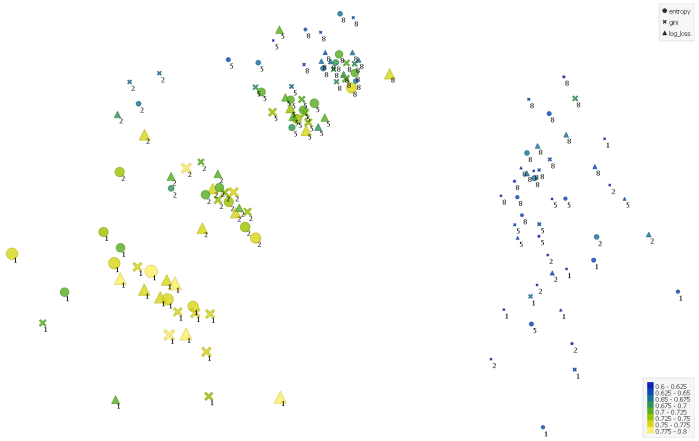


FIG. 2 – *Étude du paramétrage d'un classifieur (RF) et d'une représentation en sélectionnant une partie des résultats de classification (classifieur RF, vidéo NeutreH2, données pupillométriques, soit 136 résultats). La couleur et la taille représentent l'"accuracy". Le label texte représente le paramètre du nombre minimum d'objets pour créer une feuille dans RF. 5 groupes se distinguent : à droite (grand groupe en bleu), il s'agit du codage pupillométrique normalisé par individu (valeurs de 0 à 100%) qui donne de moins bons résultats que les données pupillométriques brutes (valeurs en mm, groupes à gauche). 4 sous-groupes se distinguent en fonction du paramètre de RF.*

Également, si l'on observe le type de normalisation utilisée dans les données pupillométriques, on trouve que les valeurs brutes en mm sont plus efficaces que les valeurs normalisées par le minimum et maximum de dilatation par enfant. Ce résultat est aussi confirmé par les experts (voir l'analyse plus locale dans la section suivante).

Ensuite nous pouvons analyser les classifieurs (voir figure 1(d) où le type de classifieurs est codé comme une couleur). Au centre du graphique, on trouve un groupe de résultats, presque tous ceux d'Adaboost. Les valeurs de probabilités produites par AdaBoost sont toutes proches de 0,5, contrairement aux autres classifieurs qui utilisent une plus grande plage de valeurs dans l'intervalle $[0, 1]$. C'est la raison pour laquelle ces résultats sont regroupés ensemble, quelle que soit la représentation utilisée. Une autre forme visuelle qui ressort est la spirale autour des deux groupes centraux. Il s'agit cette fois du classifieur 1NN. Celui-ci se distingue des autres car ses probabilités de classe sont binaires (0 ou 1, car un seul voisin est utilisé), alors que pour les autres classifieurs, ces probabilités sont plutôt des valeurs continues. Les autres résultats de KNN pour $K > 1$ se séparent également des autres car la règle du vote majoritaire dans KNN fait que souvent la probabilité de la classe atteint 0 ou 1 lorsque le vote est unanime.

3.3 Étude du paramétrage d'un classifieur et d'une représentation efficace

Notre approche permet d'étudier et de choisir des paramètres pour un classifieur. Dans les tests que nous avons réalisés, nous avons utilisé une approche de type "grid search" pour

explorer les paramètres des classifieurs. Nous prenons comme exemple le cas de RF. Pour rappel, les différents paramètres testés sont : le nombre d'arbres (10, 30, 50, 100), le critère optimisé (Gini, Entropie, "log loss"), le nombre minimum d'exemples dans une feuille (1, 2, 5, 8). Pour explorer ces résultats, nous choisissons une des représentations qui donne des valeurs d'"accuracy" les plus élevées. Il s'agit du stimulus NeutreH2, du codage pupillométrique sur données normalisées (par le minimum et maximum en mm calculés pour chaque enfant) ou brutes (valeur mesurée en mm).

La figure 2 montre les résultats les plus saillants obtenus par notre approche. On constate qu'un grand groupe de résultats à droite obtient des valeurs d'"accuracy" plus faibles que les autres. En observant ces points, on constate qu'il s'agit du codage normalisé par enfant qui est donc moins discriminant que les valeurs brutes en mm du diamètre de la pupille. Les experts du domaine apportent une explication à cette constatation : l'amplitude de la dilatation pupillaire diffère selon la classe, et la normalisation a tendance à atténuer cette différence.

Ensuite, les points à gauche du graphique se répartissent en 4 sous-groupes. En affichant sous la forme d'un label chacun des paramètres de RF, on constate que ces groupes correspondent principalement au paramètre du nombre minimum d'exemples dans une feuille. La valeur 1 est la plus efficace pour les données sélectionnées.

4 Conclusion et perspectives

Nous avons étudié dans cet article une approche visuelle pour explorer un grand nombre de résultats de classification. A partir de la prédiction des classes, il est possible de calculer une distance entre résultats, puis, avec une méthode de projection, de proposer une visualisation 2D d'un grand nombre de résultats. Dans le cadre d'une application réelle, nous avons montré quelles informations peuvent être découvertes avec cette approche, afin de mieux comprendre les représentations et les performances des classifieurs et de leur paramétrage.

Plusieurs perspectives sont envisagées. Dans le cadre d'une autre expérience sur la détection des TSA, nous allons disposer d'un plus grand nombre de résultats, ce qui nécessite l'utilisation de méthodes de projection de données plus efficaces d'un point vue calculatoire. Également, nous avons remarqué que des interactions plus fluides faciliteraient l'analyse visuelle des résultats. Enfin, une problématique complémentaire qui émerge est de comprendre pourquoi certains objets peuvent être systématiquement mal classés. Nous proposerons également des approches visuelles pour ce problème.

Références

Aguillon-Hernandez, N., Y. Mofid, M. Latinus, L. Roché, M. R. Bufo, M. Lemaire, J. Malvy, J. Martineau, C. Wardak, et F. Bonnet-Brilhault (2020). The pupil : a window on social automatic processing in autism spectrum disorder children. *Journal of Child Psychology and Psychiatry 61*(7), 768–778.

Alaiz-Rodríguez, R., N. Japkowicz, et P. Tischer (2008). Visualizing classifier performance on different domains. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, Volume 2, pp. 3–10.

Alsallakh, B., A. Hanbury, H. Hauser, S. Miksch, et A. Rauber (2014). Visual methods for analyzing probabilistic classification data. *IEEE Transactions on Visualization and Computer Graphics 20*(12), 1703–1712.

Chae, J., S. Gao, A. Ramanathan, C. A. Steed, et G. Tourassi (2017). Visualization for classification in deep neural networks. In *Proceedings of the Workshop on Visual Analytics for Deep Learning*.

Ghosh, A., P. Chaudhuri, et C. Murthy (2005). On visualization and aggregation of nearest neighbor classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*(10), 1592–1602.

Guimard-Brunault, M., N. Hernandez, L. Roché, S. Roux, C. Barthélémy, J. Martineau, F. Bonnet-Brilhault, et al. (2013). Back to basic : do children with autism spontaneously look at screen displaying a face or an object ? *Autism research and treatment 2013*.

Heyen, F., T. Munz, M. Neumann, D. Ortega, N. T. Vu, D. Weiskopf, et M. Sedlmair (2020). Clavis : An interactive visual comparison system for classifiers. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA. Association for Computing Machinery.

Luque, A., M. Mazzoleni, A. Carrasco, et A. Ferramosca (2022). Visualizing classification results : Confusion star and confusion gear. *IEEE Access 10*, 1659–1677.

Meng, L., S. van den Elzen, et A. Vilanova (2022). Modelwise : Interactive model comparison for model diagnosis, improvement and selection. *Computer Graphics Forum 41*(3), 97–108.

Neto, M. P. et F. V. Paulovich (2021). Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics 27*(2), 1427–1437.

Rauber, P. E., S. G. Fadel, A. X. Falcão, et A. C. Telea (2017). Visualizing the hidden activity of artificial neural networks. *IEEE Transactions on Visualization and Computer Graphics 23*(1), 101–110.

Schulz, A., A. Gisbrecht, et B. Hammer (2015). Using discriminative dimensionality reduction to visualize classifiers. *Neural Process. Lett. 42*(1), 27–54.

Zintgraf, L. M., T. S. Cohen, T. Adel, et M. Welling (2017). Visualizing deep neural network decisions : Prediction difference analysis. In *International Conference on Learning Representations*.

Summary

We propose a visual method to analyze a large number of supervised learning results, and we illustrate it on a real world application (detection of autistic disorders from eye tracking data) in which we were able to obtain more than 30,000 classification results. Each "representation-classifier-parameters" triplet is positioned in 2D with a dimension reduction algorithm (MDS or t-SNE) on the basis of a distance using the probability of the classes predicted for each data. With different settings of this visualization, we show that the user can visually observe and evaluate the effectiveness of the representations, classifiers and their settings.