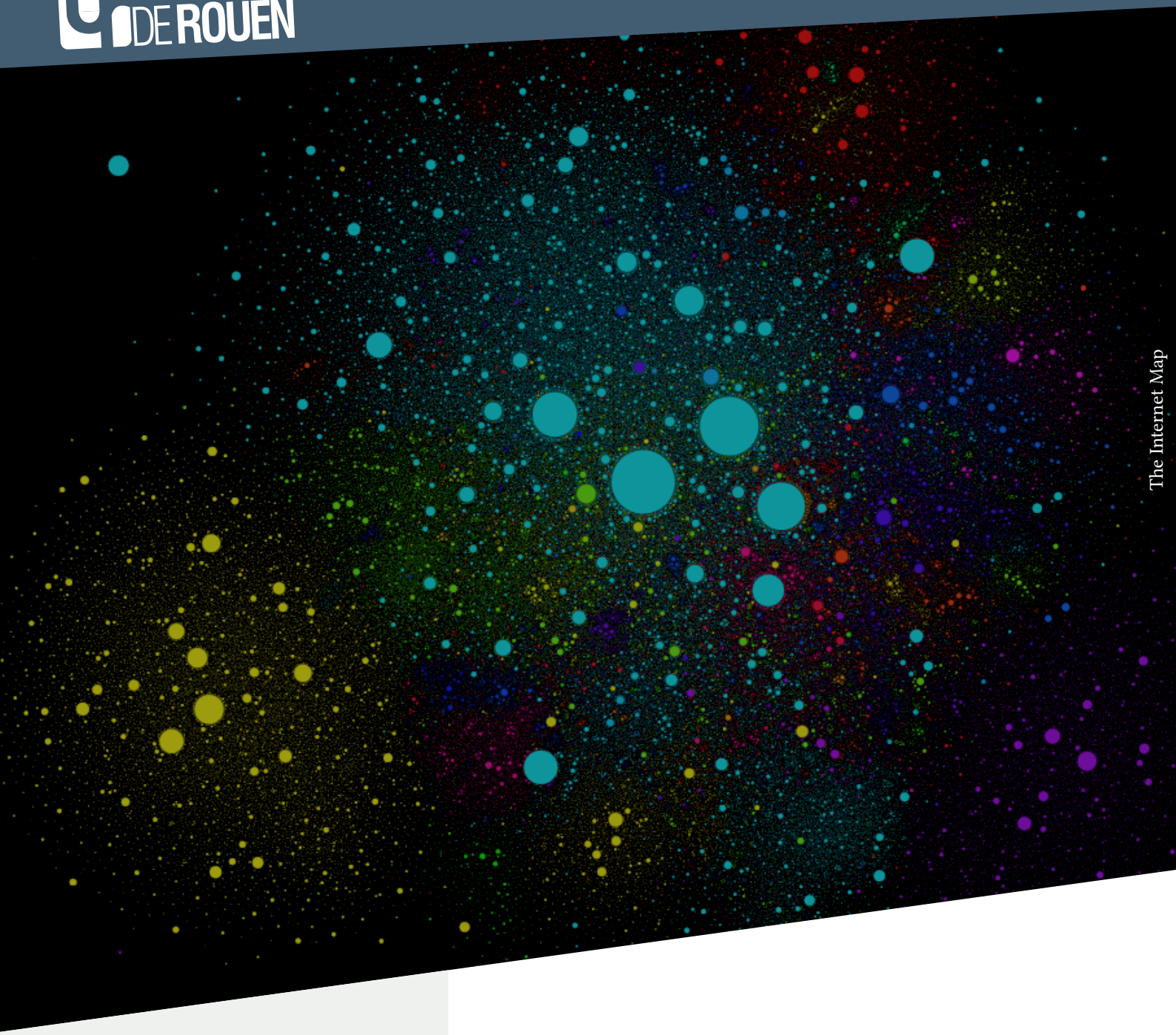
UNIVERSITÉ DE ROUEN



The Internet Map

Fusion d'Information

Année 2014-2015

Compte-rendu de TP

Mélange de classifieurs

Thomas Robert

1 | Méthode d'évaluation

Une composante importante du TP a consisté a travailler sur une fonction permettant d'évaluer les performances d'un classifieur en top i en calculant les taux de classification, rejet en ambiguité et confusion.

Ces scores sont calculés à partir d'une matrice X avec une ligne par exemple à classifier, et une colonne par classe. La valeur X_{ij} de la matrice correspond à une mesure de la confiance du classifieur dans le fait que l'exemple i est de la classe j.

Cette méthode est donc généralisable et applicable pour évaluer n'importe quel résultat, que la matrice X contienne des mesures, des rangs ou des votes.

On considère qu'un exemple i dont la classe réelle est j est "classé" en top k si X_{ij} fait parti des k plus fortes valeurs de la ligne $X_{i\bullet}$, et que la valeur X_{ij} n'a pas de valeur égale en dehors des k plus fortes valeurs de $X_{i\bullet}$, c'est à dire qu'il n'y a pas de conflit avec le score X_{ij} en dehors du top k. Notons également que X_{ij} doit être supérieur à 0, puisque les scores égaux à zéro correspondent aux cas non décidés par le classifieur.

On considère qu'un exemple i dont la classé réelle est j est "rejeté en ambiguité" en top k si X_{ij} fait parti des k plus fortes valeurs de la ligne $X_{i\bullet}$ mais que la valeur X_{ij} a au moins une valeur égale en dehors des k plus fortes valeurs de $X_{i\bullet}$, c'est à dire qu'il y a un conflit avec le score X_{ij} en dehors du top k. L'exemple i peut également être rejeté si aucun score de la ligne $X_{i\bullet}$ est supérieur à 0, c'est à dire que l'on rejette en ambiguité un exemple pour lequel le classifieur ne donne aucun résultat. Ce cas arrive en combinaison de mesure par produit par exemple.

On considère qu'un exemple i dont la classé réelle est j est "confus" en top k si X_{ij} ne fait pas parti des k plus fortes valeurs de la ligne $X_{i\bullet}$. C'est à dire les cas qui ne sont ni "classés", ni "rejetés".

2 | Performance reco Top1 et Top5

On mesure les performances des divers classifieurs en top 1 et top 5. On constate que les performances des classifieurs sont très variables, allant pour le top 1 et 60 à 90%.

En toute logique, les performances en top 5 sont supérieures à celles en top 1, allant de 88 à 96%. Voir les figures 1 et 2.

	Classif T1	Rejet T1	Classif T2	Rejet T2	Classif T3	Rejet T3	Classif T4	Rejet T4	Classif T5	Rejet T5
CI1	0.5997	4.0000e-04	0.6737	0.0017	0.7402	0.0078	0.8075	0.0171	0.8846	0
CI2	0.6999	3.0000e-04	0.7491	7.0000e-04	0.7942	0.0038	0.8374	0.0103	0.8909	0
CI3	0.7499	2.0000e-04	0.7871	0.0011	0.8226	0.0024	0.8517	0.0078	0.8891	0
CI4	0.7990	0.0011	0.8433	0.0013	0.8836	0.0060	0.9211	0.0110	0.9699	0
CI5	0.8997	3.0000e-04	0.9140	2.0000e-04	0.9295	9.0000e-04	0.9418	0.0037	0.9600	0

Figure 1 – Performances en top 1 à 5 pour les classifieurs seuls

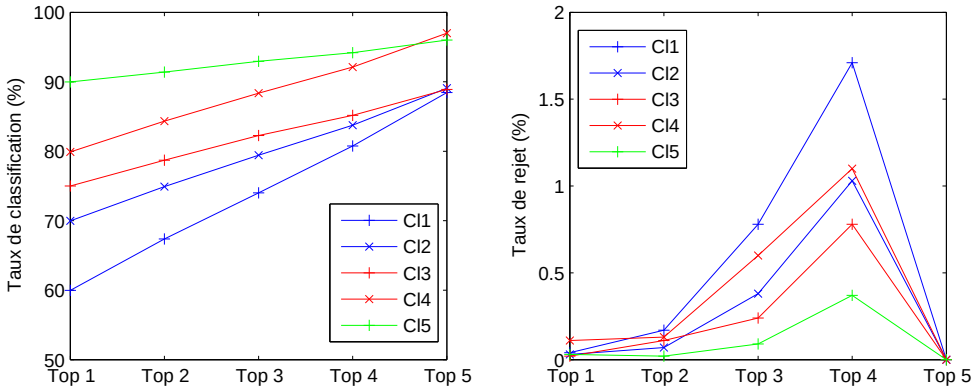


Figure 2 – Performances en top 1 à 5 pour les classifieurs seuls

3 | Méthodes de combinaison de type "Classe"

On trace pour chaque méthode (vote à la pluralité, la majorité, à la pluralité pondérée) et pour chaque jeu de données (apprentissage et test) les scores en classification, rejet d'ambiguité et confusion. Voir la figure 3.

De manière générale, les performances obtenues sont bonnes, bien meilleures que les performances des classifieurs pris séparément, montrant bien (au moins dans ce cas) l'apport des mélanges de classifieurs.

On constate sans étonnement que le vote à la pluralité provoque moins de rejet que le vote à la majorité, puisque les résultats du vote à la majorité sont les mêmes que ceux du vote à la pluralité en rejetant les cas qui ne sont pas votés par au moins 50% des classifieurs.

Sur le même principe, il n'est pas étonnant de constater que le vote à la majorité à beaucoup moins de confusion que le vote à la pluralité, puisque les cas ambigus où les classifieurs sont très partagés seront rejetés en ambiguité.

Enfin, le vote à la pluralité pondérée est celui qui offre les meilleurs résultats en classifications. Cependant, ceci peut être en partie expliqué par le fait que l'utilisation de pondération des votes fait qu'il n'y a aucun cas dans le jeu de données où il y a ambiguité. C'est donc la solution qui a le meilleur score en classification, mais qui a également le plus mauvais score (le plus fort) en confusion.

Il est donc impossible de juger qu'une de ces solutions est meilleure que les autres puisque pour chaque, une augmentation des performances en classification entrainer une augmentation du taux de confusion. Le choix de la "meilleure" solution pour un cas donné pourrait être fait en attribuant des coûts aux 3 cas (classification, rejet, confusion) par exemple.

4 | Méthodes de combinaison de type "Rang"

On essaye maintenant des méthodes de combinaison de type rang. On ne considère donc plus les probabilités en sortie des classifieurs mais simplement l'ordre de ces probabilités, correspondantes au rang de chaque prédiction.

	% Classification	% Rejet	% Confusion
Pluralité app	96.4500	2.8400	0.7100
Pluralité test	96.4700	2.7800	0.7500
Majorité app	90.4400	9.4400	0.1200
Majorité test	90.5100	9.3100	0.1800
Pondération app	98.2600	0	1.7400
Pondération test	98.3500	0	1.6500

Figure 3 – Performances en top 1 pour les méthodes de type classe

Voir les figures 4, 5 et 6.

On constate que les méthodes de type Borda-Count sont toutes très proches les unes des autres, entre 96 et 98% de bonne classification en top 1.

La meilleure des méthodes de Borda-Count est sans conteste la méthode de Borda-Count avec poids, pondérée. Ce méthode associe a chaque rang un poids qui est $c^(r-1)$ où c est une constante dans $[0, 1]$ et r le rang.

La méthode du meilleur rang peut également être intéressante si on tolère un très fort taux de rejet en ambiguité. En effet, cette méthode produit très peu d'erreurs en top 1 (0,03% en test), mais rejette beaucoup (77% en test).

	% Classification	% Rejet	% Confusion
BC moyenne app	96.1700	1.2800	2.5500
BC moyenne test	95.9700	1.1600	2.8700
BC poids app	97.7700	0.1100	2.1200
BC poids test	97.6600	0.0900	2.2500
BC moyenne pondéré app	97.0800	0	2.9200
BC moyenne pondéré test	97.1300	0	2.8700
BC poids pondéré app	98.0400	0	1.9600
BC poids pondéré test	98.1100	0	1.8900
Meilleur rang app	23.1300	76.8000	0.0700
Meilleur rang test	22.5300	77.4400	0.0300

Figure 4 – Performances en top 1 pour les méthodes de type rang

5 | Méthode de combinaison de type "Mesure"

Essayons maintenant des méthodes de combinaison de type mesure. On utilise donc directement les scores en sortie des classifieurs, affectés à chaque classe pour chaque exemple.

Ces scores peuvent être combinés par somme ou produit, pondérés ou non.

Voir les figures 7, 8 et 9.

	% Classification	% Rejet	% Confusion
BC moyenne app	99.9000	0.0500	0.0500
BC moyenne test	99.8600	0.0200	0.1200
BC poids app	99.9300	0.0400	0.0300
BC poids test	99.9300	0.0100	0.0600
BC moyenne pondéré app	99.9300	0	0.0700
BC moyenne pondéré test	99.9200	0	0.0800
BC poids pondéré app	99.9600	0	0.0400
BC poids pondéré test	99.9600	0	0.0400
Meilleur rang app	99.9300	0.0500	0.0200
Meilleur rang test	99.9700	0.0100	0.0200

Figure 5 – Performances en top 5 pour les méthodes de type rang

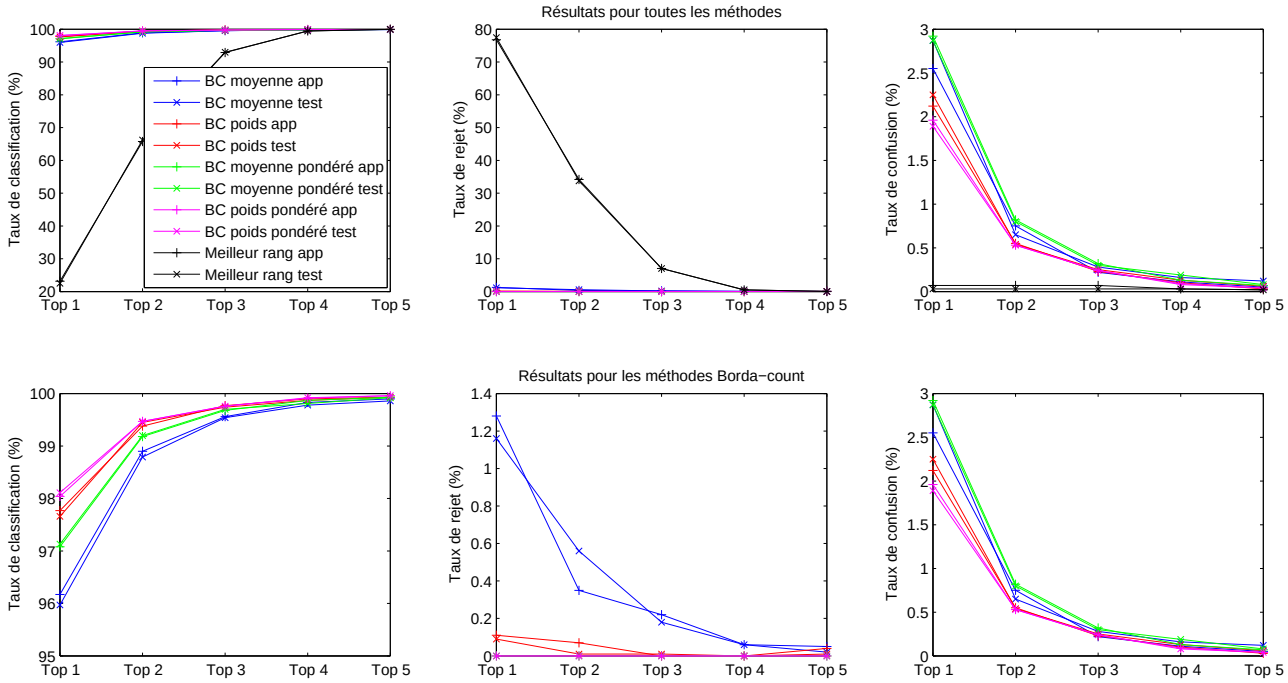


Figure 6 – Performances en top 1 à 5 pour les méthodes de type rang

Dans notre cas, les méthodes de somme donnent des meilleurs résultats que le produit sur tous les plans : taux plus fort en classification et plus faible en rejet et en confusion.

6 | Comparaison des méthodes

On se propose finalement de comparer les performances des différentes méthodes en test (les résultats en apprentissage et en test étant quasiment identique, inutile de doubler la quantité de données à analyser).

En top 1, on affiche un tableau des résultats, trié par taux de classification. Voir la figure 10. On voit que la majorité des méthodes sont proches les unes des autres, mais que le vote pondéré donne les meilleurs résultats.

	% Classification	% Rejet	% Confusion
Somme app	97.0300	0	2.9700
Somme test	97.0900	0	2.9100
Produit app	65.2100	31.4700	3.3200
Produit test	66.2600	30.3400	3.4000
Somme pondéré app	97.3400	0	2.6600
Somme pondéré test	97.5800	0	2.4200
Produit pondéré app	65.2400	31.4700	3.2900
Produit pondéré test	66.3200	30.3400	3.3400

Figure 7 – Performances en top 1 pour les méthodes de type mesure

	% Classification	% Rejet	% Confusion
Somme app	99.9500	0	0.0500
Somme test	99.9300	0	0.0700
Produit app	65.3100	31.4700	3.2200
Produit test	66.4700	30.3400	3.1900
Somme pondéré app	99.9600	0	0.0400
Somme pondéré test	99.9500	0	0.0500
Produit pondéré app	65.3100	31.4700	3.2200
Produit pondéré test	66.4700	30.3400	3.1900

Figure 8 – Performances en top 5 pour les méthodes de type mesure

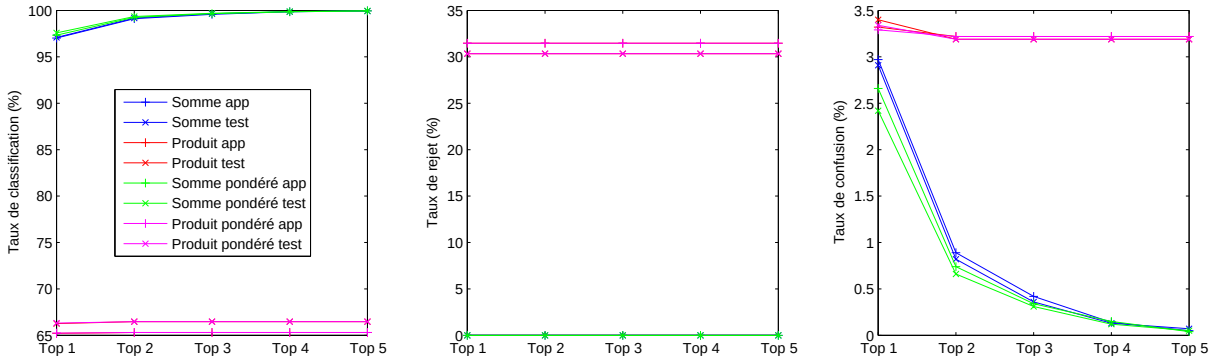


Figure 9 – Performances en top 1 à 5 pour les méthodes de type mesure

Il peut être intéressant de chercher le front de Pareto des solutions à notre disposition, afin de savoir quelles sont réellement les solutions les plus intéressantes au sens d'une optimisation multi-critère visant a maximiser le taux de classification et minimiser le taux de confusion (mathématique, cela revient également à minimiser le taux de rejet). Voir la figure 11.

On constate que le front de Pareto en top 1 contient le vote à la pluralité, le vote à la majorité, le vote pondéré et le meilleur rang. Ces résultats sont particulièrement étonnant puisqu'ils ne font apparaitre quasiment que des méthodes de type vote, et une méthode de type rang qui rejette énormément lui permettant d'avoir un taux imbattablement faible en confusion la plaçant dans le front.

	% Classification	% Rejet	% Confusion
Pondération test	98.3500	0	1.6500
BC poids pondéré test	98.1100	0	1.8900
BC poids test	97.6600	0.0900	2.2500
Somme pondéré test	97.5800	0	2.4200
BC moyenne pondéré test	97.1300	0	2.8700
Somme test	97.0900	0	2.9100
Pluralité test	96.4700	2.7800	0.7500
BC moyenne test	95.9700	1.1600	2.8700
Majorité test	90.5100	9.3100	0.1800
Produit pondéré test	66.3200	30.3400	3.3400
Produit test	66.2600	30.3400	3.4000
Meilleur rang test	22.5300	77.4400	0.0300

FIGURE 10 – Performances en top 1 pour toutes les méthodes

	% Classification	% Rejet	% Confusion
Pluralité test	96.4700	2.7800	0.7500
Majorité test	90.5100	9.3100	0.1800
Pondération test	98.3500	0	1.6500
Meilleur rang test	22.5300	77.4400	0.0300

FIGURE 11 – Performances en top 1 du front de Pareto

Cependant, on ne peut bien sûr pas généraliser ces résultats obtenus sur un cas particulier. Par ailleurs, il est important de noter que les différences entres les méthodes sont très faibles pour la majorité d'entre elles et que ce classement est donc peu significatif.

Enfin, on peut également regarder l'évolution des performances des différentes méthodes du top 1 au top 5 (affiché uniquement pour les méthodes ayant un taux de classification supérieur à 95% afin que le graphe reste lisible). Voir la figure 12. Globalement, les résultats restent très "parallèle", une méthode dépasse rarement une autre.

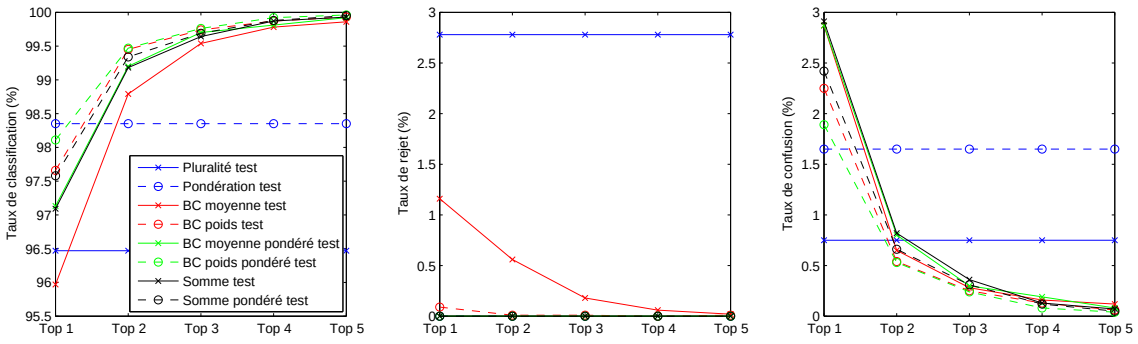


FIGURE 12 – Performances en top 1 à 5 d'une partie des méthodes