Les histogrammes quasi-continus

Olivier Strauss, Frédéric Comby

LIRMM, Université Montpellier II, 161, rue Ada,
34392 Montpellier Cedex 5, France
e-mail : strauss@lirmm.fr, comby@lirmm.fr

RÉSUMÉ : *Nous présentons, dans cet article, un outil permettant d'estimer la probabilité imprécise d'un événement réel imprécis basé sur l'accumulation d'événements réels imprécis.*

ABSTRACT: *This paper describes a new method, called rough histograms, that performs an imprecise estimation of the probability of an imprecise event based on imprecise real data.*

MOTS CLÉS : *histogramme, probabilité imprécise, possibilité, ensembles flous grossiers.*

KEYWORDS: *Histograms, imprecise probabilities, possibility, fuzzy rough sets.*

1. Introduction.

Le concept de densité de probabilité est fondamental en statistique. Lorsqu'on connaît la densité de probabilité $f(x)$ d'un événement réel x, il est possible de calculer la probabilité de n'importe quel événement imprécis donné sous forme d'intervalle : [a,b] grâce à : $P([a, b]) = \int_a^b f(x).dx$.

Lorsque cette densité est inconnue, il est possible d'en faire une estimation en utilisant un échantillon de n observations réelles ($x_1 \ldots x_n$). L'approche paramétrique suppose connue la forme de la densité de probabilité et consiste à en estimer les paramètres. Depuis les travaux de Parzen (Parzen, 1979), la communauté scientifique s'intéresse de plus en plus aux estimateurs non-paramétriques. La démarche non-paramétrique est moins rigide. Elle suppose bien sûr qu'il existe une densité de probabilité sous-jacente mais laisse les données s'exprimer plutôt que de les contraindre à adopter une forme paramétrique pré-déterminée.

Lorsque les données réelles permettant l'estimation de la densité de probabilité sont imprécises, cette imprécision devrait se répercuter sur la densité de probabilité estimée pour donner une sorte de "densité imprécise de probabilité" (Walley, 1991). A notre connaissance, aucune méthode d'estimation de densité ne prend en compte la précision des données d'apprentissage.

Nous proposons, dans cet article, une nouvelle technique permettant d'estimer une probabilité imprécise à partir d'un échantillon d'événements imprécis en substituant à la notion de densité de probabilité, celle d'histogramme quasi-continu.

2. Les histogramme quasi-continus.

La théorie des histogrammes quasi-continus (HQC) est basée sur celle des sous ensembles flous grossiers (Dubois et al., 1990). Le but est de dissocier la granularité de l'histogramme de la précision des informations que l'on en extrait.

2.1 Partition floue de l'espace.

Pour atténuer les effets dus au caractère arbitraire du partitionnement de l'espace, les histogrammes quasi-continus sont construits sur une partition floue de l'espace réel (Bezdec 1981). Une partition binaire, de granularité Δ, d'un intervalle Ψ de l'univers du discours Ω est un ensemble d'intervalles disjoints de largeur Δ recouvrant Ψ. Une partition floue de Ψ ayant la même granularité est un ensemble d'intervalles flous C_k de fonction d'appartenance $\mu_k(x)$ dont la largeur ($\int \mu_k(x)dx$) est égale à Δ.

2.2 Accumulation de données précises.

Soit $(x_1 \ldots x_n)$, un échantillon de n données. L'accumulation dans l'histogramme quasi-continu défini par la partition des p intervalles flous $(C_1 \ldots C_p)$ est donnée par :

$$Acc_k = \sum_{i=1}^{n} \mu_k(x_i) \qquad [1]$$

où Acc_k, l'accumulateur associé à la cellule C_k, représente le nombre des données de l'échantillon compatibles avec l'intervalle C_k de fonction d'appartenance μ_k.

2.3 Accumulation de données imprécises.

Un échantillon de données imprécises peut être aisément représenté par un ensemble d'intervalles ordinaires ou flous $(X_1 \ldots X_n)$. On doit alors remplacer l'accumulateur précis de la formule [1] par un accumulateur imprécis dont les bornes supérieures ($\overline{Acc_k}$) et inférieures ($\underline{Acc_k}$) sont estimées comme suit :

$$\overline{Acc_k} = \sum_{i=1}^{n} \Pi(C_k;X_i) \qquad \underline{Acc_k} = \sum_{i=1}^{n} N(C_k;X_i) \qquad [2]$$

où $\Pi(C_k;X_i)$ (rsp. $N(C_k;X_i)$) est la possibilité (rsp. la nécessité) de la donnée X_i restreinte à la cellule C_k. Cette imprécision de l'accumulateur caractérise l'adéquation entre la granularité de la partition sur laquelle est construit l'histogramme et la précision des données accumulées.

3. Probabilité d'un événement.

3.1 Probabilité précise.

Soit $(x_1 \ldots x_n)$, un échantillon de n données issues d'une distribution de probabilité de densité p(x), et W_Γ=[m-Γ/2,m+Γ/2] un intervalle de granularité Γ. Il existe entre $P(W_\Gamma)$ et p(x) la relation suivante :

$$P(W_\Gamma) = \int_{m-\Gamma/2}^{m+\Gamma/2} p(x)d(x) = \lim_{n\to\infty} \frac{Nb(W_\Gamma;(x_i))}{n} \quad [3]$$

où $Nb(W_\Gamma;(x_i))$ est le nombre de données de (x_i) appartenant à l'intervalle W_Γ et $P(W_\Gamma)$ la probabilité d'avoir un élément de la distribution (x_i) dans l'intervalle W_Γ. La densité de probabilité en m (centre de l'intervalle W_Γ) peut être vue comme une limite de $P(W_\Gamma)$ lorsque Γ tend vers 0. Le cardinal de l'échantillon considéré étant fini, on peut considérer que la distribution est uniforme dans l'intervalle (Silvermann 1998) et l'on affecte au centre de l'intervalle la densité empirique de probabilité $\hat{p}(m)$.

$$\hat{p}(m) \cong \frac{1}{\Gamma}\frac{Nb(W_\Gamma;(x_i))}{N} \quad [4]$$

Lorsque l'échantillon des données est fini, la notion générale de densité de probabilité doit être remplacée par celle de densité de probabilité à granularité Γ pour prendre en compte l'effet de la densité locale des données sur l'estimation.

3.2 Probabilité imprécise.

Lorsque les données sont imprécises, l'estimation de la probabilité est imprécise. Soit $(X_1 \ldots X_n)$ un échantillon de n données imprécises, et l'ensemble W_Γ de granularité Γ. À la notion de dénombrement des données appartenant à l'intervalle W_Γ on doit substituer celle de nombre supérieur et inférieur de données compatibles avec W_Γ :

$$\overline{Nb}(W_\Gamma;(x_i)) = \sum_{i=1}^{n} \Pi(W_\Gamma;X_i) \quad \text{et } \underline{Nb}(W_\Gamma;(x_i)) = \sum_{i=1}^{n} N(W_\Gamma;X_i) \quad [5]$$

On représente alors la probabilité empirique $P(W_\Gamma)$ par ses valeurs extrêmes :

$$\overline{P}(W_\Gamma) = \frac{\overline{Nb}(W_\Gamma;(x_i))}{\overline{Nb}(W_\Gamma;(x_i)) + \underline{Nb}(W_\Gamma^c;(x_i))}; \underline{P}(W_\Gamma) = \frac{\underline{Nb}(W_\Gamma;(x_i))}{\underline{Nb}(W_\Gamma;(x_i)) + \overline{Nb}(W_\Gamma^c;(x_i))} \quad [6]$$

avec W_Γ^c complémentaire de W_Γ. Il est facile de vérifier que $\overline{P}(W_\Gamma)$ et $\underline{P}(W_\Gamma)$ sont bien les bornes supérieure et inférieure de toutes les probabilités que l'on peut construire sur des dénombrements bâtis sur l'échantillon imprécis $(X_1 \ldots X_n)$.

4. Estimation de la probabilité imprécise d'un événement imprécis.

La manipulation des formules [5] et [6] est particulièrement malaisée dans le cas de calculs récursifs ou itératifs (e.g. recherche de fractile ou de modes imprécis). Cependant, si on connaît la valeur des accumulateurs d'un HQC de granularité Δ sur un support Ψ, il est possible de faire une estimation du nombre de données X_i compatibles avec tout ensemble de granularité $\Gamma > \Delta$ dont le support est inclus dans Ψ. Les données étant imprécises, les estimations de ces nombres sont imprécises.

Nous ne donnons ici que les formules dans le cas imprécis. Le cas précis peut être facilement déduit du cas imprécis en posant $\overline{Acc_k} = \underline{Acc_k} = Acc$.

Le pire cas est obtenu en utilisant un transfert possibiliste des accumulateurs :

$$\underline{Nb}(W_\Gamma;(x_i)) = \sum_{k=0}^{p} \underline{Acc_k}\Pi(C_k;W_\Gamma) \text{ et } \overline{Nb}(W_\Gamma;(x_i)) = \sum_{k=0}^{p} \overline{Acc_k}N(C_k;W_\Gamma) \quad [7]$$

Le cas “le plus probable” est obtenu en utilisant une technique inspirée du transfert pignistique de croyance suggéré par Philippe Smets (Smets, 1994) :

$$\underline{Nb}(W_\Gamma;(x_i)) = \sum_{k=0}^{p} \underline{Acc_k}\frac{|W_\Gamma \cap C_k|}{|C_k|} \text{ et } \overline{Nb}(W_\Gamma;(x_i)) = \sum_{k=0}^{p} \overline{Acc_k}\frac{|W_\Gamma \cap C_k|}{|C_k|} \quad [8]$$

On utilise alors les formules [3], [4] et [6] pour retrouver la probabilité de W_Γ.

5. Conclusion et discussion.

Nous avons montré, dans cet article, qu’il était possible d’utiliser un histogramme quasi-continu pour estimer la probabilité imprécise d’un événement de type intervalle à partir d’un échantillon de données imprécises. Cet outil est simple à utiliser. Dans d’autres articles nous avons montré qu’il permettait de réaliser des statistiques d’ordre (Strauss et al., 1999), des statistiques modales uni- et multivariées (Strauss et al., 2000) et nous l’avons appliqué à l’estimation du mouvement dominant dans une séquence d’images (Comby et al., 2001).

6. Références

Bezdek J., *Pattern recognition with fuzzy objective function algorithms*. N.Y., Plenum, 1981.

Dubois D., Prade H., Rough fuzzy sets and fuzzy rough sets. *International journal of general systems*, 17(2-3), pp191-200, 1990.

Comby F., Strauss O., Aldon M.J., Possibility theory and rough histograms for motion estimation in a video sequence, *IWVF4*, pp. 473-483. Capri, Italy, 2001.

Parzen, E., Nonparametric statistical data modeling, J*ournal of the American Statistical Association*, vol. 74, pp. 105-131, 1979.

Silverman B., *Density estimation for statistics and data analysis*, Chapman and Hall, 1998.

Smets P., The transferable belief model. *Artificial Intelligence* (66) pp. 191-243, 1994

Strauss O., Lavarec E., Histogrammes approchés : application aux statistiques d’ordre, *LFA99*, pp .161-168, Valenciennes, France, Octobre 1999.

Strauss O., Comby F., Aldon M.J., Rough histograms for robust statistics. *International Conference on Pattern Recognition*, vol2, pp. 688-691. September 2000.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall, New York, 1991.