Ανάλυση ιστογράμματος και ομαδοποίηση δεδομένων

Γεώργιος Τζιρίτας
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Φεβρουάριος 2019

Με την ανάλυση του ιστογράμματος η τμηματοποίηση καθίσταται ένα πρόβλημα κατάταξης των σημείων της εικόνας σε κλάσεις. Από το ιστόγραμμα, μετά από τη λείανσή του, εφόσον είναι αναγκαία, προσδιορίζονται οι επικρατούσες τιμές της ολικής κατανομής. Η κατάταξη των σημείων της εικόνας γίνεται με βάση αυτές τις τιμές, ή ενδεχόμενα μέσω της κατανομής ανά κλάση, εφόσον θα μπορούσε να υπολογισθεί αξιόπιστα. Η μέθοδος αυτή δεν εξασφαλίζει οπωσδήποτε τη συνεκτικότητα των αντικειμένων που εντοπίζονται, αφού η κατάταξη κάθε σημείου βασίζεται αποκλειστικά στη φωτεινή ένταση του δοσμένου μόνο σημείου. Η αξιοπιστία της μεθόδου εξαρτάται από την ομοιογένεια του φωτισμού και από το μέγεθος των αντικειμένων, που επιπλέον ως προς το είδος θα πρέπει να είναι ολίγα τον αριθμό.

1 Μοντέλο κατανομής πιθανότητας

Ο χωρισμός του ιστογράμματος σε διαστήματα μπορεί να βασισθεί σε κάποιο αρχικό μοντέλο για τη συνάρτηση πυκνότητας πιθανότητας της μεταβλητής που εκφράζει τη φωτεινή ένταση. Ας θεωρήσουμε την περίπτωση δύο επικρατουσών τιμών, ή ενός είδους αντικειμένου στο προσκήνιο που ανιχνεύεται σε αντίθεση προς το παρασκήνιο της εικόνας. Ας υποθέσουμε ότι η τυχαία μεταβλητή του μοντέλου ακολουθεί και στις δύο περιπτώσεις την κανονική κατανομή, με την ίδια διασπορά, σ^2, και μέση τιμή μ_0 για το βάθος της εικόνας, και μ_1 για το αντικείμενο. Με το κριτήριο της μέγιστης αληθοφάνειας το κατώφλι ανίχνευσης τοποθετείται στη μέση της απόστασης μεταξύ των θέσεων των δύο τοπικά μεγίστων τιμών του ιστογράμματος

$$\kappa = \frac{\mu_0 + \mu_1}{2}. \tag{1}$$

Αν επιπλέον θεωρήσουμε ότι είναι γνωστές οι *a priori* πιθανότητες του αντικειμένου (P_1) και συμπληρωματικά του βάθους της εικόνας ($P_0 = 1 - P_1$), τότε η μεγιστοποίηση της *a posteriori* πιθανότητας δίδει το ακόλουθο κατώφλι για την ανίχνευση του αντικειμένου

$$\kappa = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{P_0}{P_1}. \tag{2}$$

Πρόκειται για τη λύση της εξίσωσης

$$P_0 e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} = P_1 e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}.$$

Ο προσδιορισμός των παραμέτρων $\mu_0, \mu_1, \sigma^2, P_0, P_1$ μπορεί να γίνει με τη βοήθεια του ιστογράμματος, με ανάλυση της μίξης των δύο κατανομών.

2 Ομαδοποίηση δεδομένων

Ο προσδιορισμός του κατωφλιού μπορεί επίσης να βασισθεί σ΄ ένα κριτήριο αυτόματης ομαδοποίησης. Αν μείνουμε στην περίπτωση των δύο επικρατουσών τιμών, ένα τέτοιο κριτήριο προς ελαχιστοποίηση είναι το ακόλουθο

$$S_I^2 = \sum_{i=0}^{k-1} p_i(i-\mu_0)^2 + \sum_{i=k}^{N-1} p_i(i-\mu_1)^2, \tag{3}$$

όπου p_i είναι η συχνότητα εμφάνισης, ή εμπειρική πιθανότητα, της τιμής i, για ένα σύνολο από N δυνατές τιμές. Ζητούνται οι τιμές του κατωφλιού k, και των παραμέτρων μ_0 και μ_1 που ελαχιστοποιούν το παραπάνω κριτήριο. Οι τιμές αυτές μπορούν να προσδιορισθούν χρησιμοποιώντας δύο αναγκαίες συνθήκες που ισχύουν στη θέση του ελάχιστου του κριτηρίου. Η πρώτη κατηγορία αναγκαίων συνθηκών προκύπτει για δοσμένο κατώφλι, και δίδει τις αντιπροσωπευτικές τιμές των δύο κλάσεων

$$\mu_0 = \frac{\sum_{i=0}^{k-1} ip_i}{\sum_{i=0}^{k-1} p_i} \quad \text{και} \quad \mu_1 = \frac{\sum_{i=k}^{N-1} ip_i}{\sum_{i=k}^{N-1} p_i} \tag{4}$$

Η δεύτερη αναγκαία συνθήκη δίδει το κατώφλι για δοσμένες αντιπροσωπευτικές τιμές

$$k = \lceil \frac{\mu_0 + \mu_1}{2} \rceil \tag{5}$$

Η διαδοχική χρήση των παραπάνω συνθηκών σ΄ ένα επαναληπτικό αλγόριθμο, επιτρέπει την ελαχιστοποίηση του κριτηρίου και τον προσδιορισμό του κατωφλιού που οδηγεί στην τμηματοποίηση της εικόνας.

Το κριτήριο της Εξίσωσης (3) εκφράζει την ελαχιστοποίηση της διασποράς των τιμών εντός των ομάδων. Ισοδύναμα μπορεί να απαιτηθεί η μεγιστοποίηση της απόστασης μεταξύ των ομάδων που δίδεται από την ακόλουθη σχέση

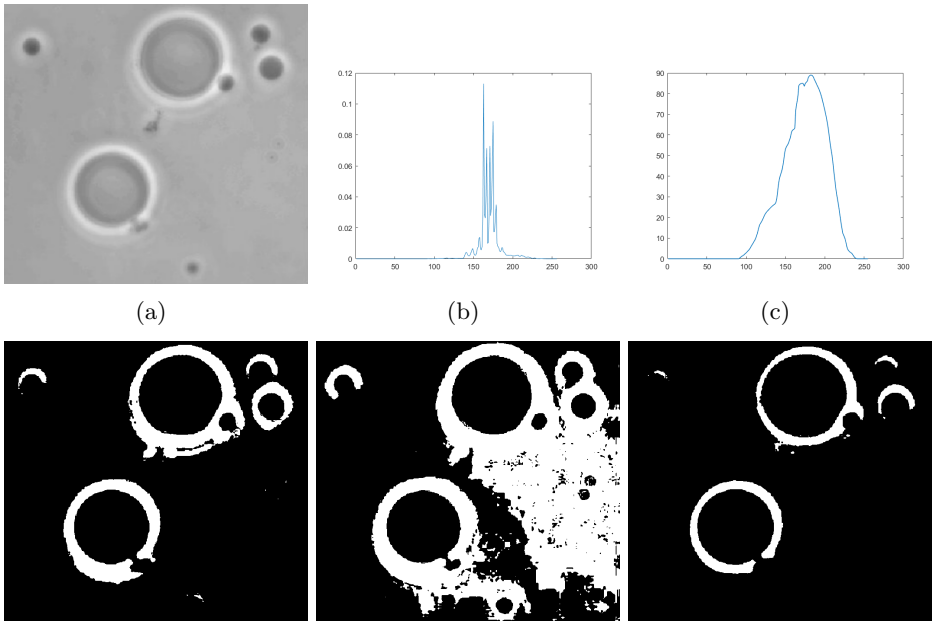
$$S_B^2 = P_0 P_1(\mu_1 - \mu_0)^2, \tag{6}$$

όπου $P_0 = \sum_{i=0}^{k-1} p_i$ και $P_1 = \sum_{i=k}^{N-1} p_i$. Η εύρεση της συνθήκης μεγιστοποίησης της μεταξύ των ομάδων απόστασης μπορεί να γίνει με διεξοδική διερεύνηση των δυνατών διαχωριστικών τιμών k. Η μέθοδος αυτή προτάθηκε από τον Otsu [3].

Η τμηματοποίηση που επιτυγχάνεται με την ομαδοποίηση των δεδομένων είναι τόσο περισσότερο ακριβής, όσο ο αριθμός των κλάσεων ανταποκρίνεται καλύτερα στα δεδομένα. Σε αυτή την περίπτωση το κριτήριο της Εξίσωσης (6) με μεγάλη πιθανότητα παρουσιάζει ένα μοναδικό μέγιστο. Τότε και ο επαναληπτικός αλγόριθμος αναμένεται να συγκλίνει στη μοναδική βέλτιστη λύση. Στο Σχήμα 1 δίδονται αποτελέσματα σε μια εικόνα προερχόμενη από μικροσκόπιο, όπου η ομαδοποίηση σε τρεις κλάσεις φαίνεται να είναι περισσότερο αντίστοιχη των τιμών της εικόνας. Ωστόσο η μέγιστη τιμή του κριτηρίου S_B^2 διαχωρίζει με ικανοποιητικό τρόπο σε δύο κλάσεις. Το κριτήριο απόστασης της Εξίσωσης (6) παρουσιάζει δύο τοπικά μέγιστα (Σχήμα 1(c)). Το αποτέλεσμα του επαναληπτικού αλγορίθμου εξαρτάται από την αρχικοποίηση των ομάδων. Μπορεί να είναι παρόμοιο με το βέλτιστο (Σχήμα 1(d)), αλλά μπορεί και να είναι όπως αυτό του Σχήματος 1(e)) που προκύπτει από το δεύτερο τοπικό μέγιστο.

Οι μέθοδοι ομαδοποίησης των δεδομένων μπορούν να επεκταθούν για περισσότερες των δύο τελικές αποχρώσεις. Το αποτέλεσμα στο Σχήμα 1(f) έχει προκύψει για τρεις ομάδες. Ένα άλλο παράδειγμα εφαρμογής με τρεις αποχρώσεις δίδεται στο Σχήμα 2.

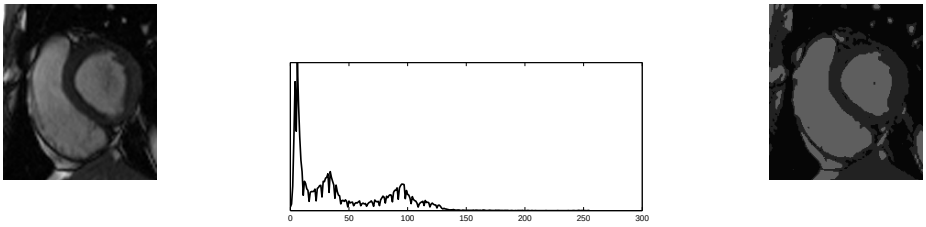
Κάποια από τα μειονεκτήματα της ανάλυσης του ιστογράμματος, όπως η ύπαρξη μικρών στην έκταση αντικειμένων ή η ανομοιογένεια του φωτισμού, που αναφέρθησαν εισαγωγικά, μπορούν να αντιμετωπισθούν με την εφαρμογή της μεθόδου τοπικά κατά τμήματα της εικόνας.



(a) (b) (c)

(d) (e) (f)

Σχήμα 1: Το ιστόγραμμα της εικόνας (a), το κριτήριο βέλτιστου διαχωρισμού (c) και αποτελέσματα τμηματοποίησης.



(a) (b) (c)

Σχήμα 2: Το ιστόγραμμα της εικόνας αριστερά και το αποτέλεσμα της τμηματοποίησης με χρήση τριών αποχρώσεων.

3 Ομαδοποίηση δεδομένων έγχρωμων ή πολυφασματικών εικόνων

Στην περίπτωση των διανυσματικών δεδομένων, όπως είναι αυτή των έγχρωμων ή των πολυφασματικών εικόνων, δεν μπορεί να εφαρμοσθεί άμεσα η ανάλυση ιστογράμματος. Σε αυτή την περίπτωση ζητείται η απευθείας εύρεση από τα δεδομένα ενός συνόλου αντιπροσωπευτικών διανυσμάτων $\{\mathbf{c}(k) : k = 1, \ldots, K\}$. Το σύνολο των αντιπροσωπευτικών χρωμάτων κατασκευάζεται από ένα σύνολο διανυσμάτων εκμάθησης $\{\mathbf{x}(n) : n = 1, \ldots, N_t\}$ και βασίζεται σ΄ ένα κριτήριο ελάχιστης απόκλισης, όπως τετραγωνικής, που ορίζεται

$$D = \sum_{k=1}^{K} \sum_{\mathbf{x}(n) \in \mathcal{S}_k} \|\mathbf{x}(n) - \mathbf{c}(k)\|^2 \tag{7}$$

για K κλάσεις $\mathcal{S}_k$. Η τετραγωνική απόκλιση εκφράζει τη διασπορά των διανυσμάτων εντός των ομάδων/κλάσεων.

Η βέλτιστη ομαδοποίηση ικανοποιεί δύο αναγκαίες συνθήκες για την ελαχιστοποίηση της D. Για δοσμένη κλάση ο καλύτερος αντιπρόσωπος είναι το κέντρο βάρους

$$\mathbf{c}(k) = \frac{1}{\mathrm{card}[\mathcal{S}_k]} \sum_{\mathbf{x}(n) \in \mathcal{S}_k} \mathbf{x}(n) \tag{8}$$

Για δοσμένο σύνολο αντιπροσωπευτικών χρωμάτων η καλύτερη κατάταξη ενός διανύσματος x συνίσταται στην επιλογή του πλησιέστερου αντιπρόσωπου

$$\|\mathbf{x} - \mathbf{c}(k)\| < \|\mathbf{x} - \mathbf{c}(l)\| \;\; \forall l \neq k \quad \Rightarrow \quad \mathbf{x} \in \mathcal{S}_k \tag{9}$$

Η χρησιμοποίηση των δύο αυτών αναγκαίων συνθηκών δίδει έναν επαναληπτικό αλγόριθμο κατασκευής ενός συνόλου αντιπροσωπευτικών χρωμάτων. Ο αλγόριθμος είναι γνωστός με το όνομα *k-means*.

- Αρχικό βήμα: Αρχικά αντιπροσωπευτικά χρώματα, $i = 1$, και αρχική μεγάλη τιμή για την απόκλιση $D^{(0)}$
- Βήμα 1: Εύρεση των κλάσεων (Εξίσωση (9))
- Βήμα 2: Υπολογισμός της απόκλισης $D^{(i)}$
- Βήμα 3: Έλεγχος σύγκλισης

  $$\text{Αν} \;\; \frac{D^{(i-1)} - D^{(i)}}{D^{(i-1)}} \leq \epsilon, \quad \text{τέλος}$$

  Διαφορετικά, προσαύξηση του i, και συνέχιση των επαναλήψεων
- Βήμα 4: Εύρεση του καλύτερου αντιπρόσωπου για κάθε κλάση (Εξίσωση (8)), κι επιστροφή στο Βήμα 1.

Ο επαναληπτικός αλγόριθμος συγκλίνει σε ένα τοπικό ελάχιστο του κριτηρίου ελάχιστης απόκλισης, που εξαρτάται από την αρχικοποίηση. Ένα αποτέλεσμα τμηματοποίησης δίδεται στο Σχήμα 3, όπου έγιναν 5 τυχαίες αρχικοποιήσεις και επιλέχθηκε το καλύτερο, με βάση το κριτήριο, αποτέλεσμα.

Ευρεία επισκόπηση μεθόδων ομαδοποίησης δεδομένων δίδεται στα [1] και [2].

(a) (b)

Σχήμα 3: Τμηματοποίηση μιας έγχρωμης εικόνας με ομαδοποίηση σε 5 κλάσεις.

4 Βιβλιογραφία

[1] A. Jain and R. Dubes, Algorithms for clustering data, Prentice Hall, 1988.

[2] A. Jain, M. Murty and P. Flynn, Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264–323, 1999.

[3] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, No 1, pp. 62–66, 1979.