ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Η κατηγοριοποίηση (classification) είναι μια διαδικασία που εντάσσεται γενικότερα στην περιοχή της επιβλεπόμενης μάθησης όπου οι κατηγορίες στις οποίες χωρίζεται αρχικά το σύνολο δεδομένων εκπαίδευσης, είναι γνωστές. Παραδείγματα εφαρμογών της κατηγοριοποίησης περιλαμβάνουν αναγνώριση προτύπων και εικόνας, ιατρική διάγνωση, έγκριση δανείων, ανίχνευση λαθών σε βιομηχανικές εφαρμογές, όπως επίσης και κατηγοριοποίηση των τάσεων στην οικονομία.

Το σύνολο δεδομένων ελέγχου χρησιμοποιείται για να αξιολογηθεί η απόδοση του μοντέλου που δημιουργείται μετά την εφαρμογή ενός αλγορίθμου κατηγοριοποίησης, αν δηλαδή τα δεδομένα ελέγχου τοποθετήθηκαν στις σωστές κλάσεις. Η απόδοση του υπό εκπαίδευση μοντέλου ελέγχεται μετρώντας το ποσοστό των σημείων που τοποθετήθηκαν στις σωστές κλάσεις. Στην περίπτωση που η ακρίβεια του μοντέλου θεωρηθεί ικανοποιητική τότε το ίδιο μοντέλο εφαρμόζεται σε ένα σύνολο από άγνωστα δεδομένα, όπου ανάλογα με την κατάταξη των δεδομένων στις αντίστοιχες κλάσεις λαμβάνονται και οι ανάλογες αποφάσεις.

Ένα παράδειγμα κατηγοριοποίησης είναι ο έλεγχος ασφάλειας ενός αεροδρομίου για τον καθορισμό των επιβατών από πιθανούς τρομοκράτες ή εγκληματίες. Για να γίνει αυτό, σαρώνεται με ειδικό σαρωτή το πρόσωπο κάθε επιβάτη και αναγνωρίζεται το βασικό του πρότυπο, όπως η απόσταση μεταξύ των ματιών, το μέγεθος και σχήμα στόματος, σχήμα κεφαλιού, κλπ. Αυτό το πρότυπο συγκρίνεται με τις καταχωρήσεις μιας βάσης δεδομένων για να διαπιστωθεί, εάν ταιριάζει με κάποια πρότυπα που συσχετίζονται με γνωστοποιημένους παραβάτες του νόμου.

Γνωρίσματα

Παρατηρήσεις

Ονοματεπώνυμο	Απόσταση ματιών	Χρώμα ματιών	Σχήμα κεφαλιού	Σχήμα φρυδιών	Παραβάτης
Audry Overbey	3.2	Γαλανά	Οβάλ	Οριζόντια	ΝΑΙ
Alec Gusler	2.1	Καφέ	Τετραγωνισμένο	Καμπυλωτά	ΟΧΙ
Ashleigh Cerezo	3.0	Καφέ	Στρογγυλό	Ανωδικά	ΝΑΙ

Γνώρισμα κλάσης

Σύνολο δεδομένων για παραβάτες

Κάθε γραμμή του συνόλου δεδομένων ανήκει σε ένα πρότυπο. Οι γραμμές καλούνται αντικείμενα, παραδείγματα ή παρατηρήσεις. Οι στήλες αναφέρονται σε μια ιδιότητα των αντικειμένων, όπως πχ η απόσταση των ματιών, το χρώμα των ματιών κλπ. Οι στήλες καλούνται και πεδία (fields), μεταβλητές (variables), γνωρίσματα (attributes) ή χαρακτηριστικά (features). Το γνώρισμα το οποίο περιέχει την απάντηση για το αν ο επιβάτης που ελέγχεται είναι εγκληματίας ή οχι, είναι το γνώρισμα της κλάσης. Στο Σχήμα 3.1 παρουσιάζεται το σύνολο δεδομένων αυτού του παραδείγματος.

Στάδια κατηγοριοποιήσης

Η κατηγοριοποίηση περιλαμβάνει τρία στάδια, το στάδιο της εκμάθησης, το στάδιο του ελέγχου του μοντέλου και το στάδιο της εφαρμογής.

- **Εκμάθηση (Learning).** Στο στάδιο αυτό, μια μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων προκειμένου να σχηματιστεί ένα μοντέλο. Η κατασκευή ή εκπαίδευση του μοντέλου καθοδηγείται από τις τιμές του γνωρίσματος της κλάσης και για τον λόγο αυτό η διαδικασία ονομάζεται επιβλεπόμενη μάθηση. Το σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση του μοντέλου, ονομάζεται σύνολο εκπαίδευσης (training data set). Η επιλογή του συνόλου εκπαίδευσης είναι καθοριστικής σημασίας, γιατί το μοντέλο που θα προκύψει θα αποτυπώνει σχέσεις που υπάρχουν στο σύνολο εκπαίδευσης.
- **Έλεγχος (Testing**). Στο στάδιο αυτό ακολουθεί η διαδικασία ελέγχου της ακρίβειας του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά. Το μοντέλο τροφοδοτείται με αντικείμενα, των οποίων η κλάση είναι γνωστή. Αναλύοντας τα στοιχεία των ανεξάρτητων γνωρισμάτων κάθε αντικειμένου, το μοντέλο προβλέπει την κλάση του αντικειμένου και στη συνέχεια το συγκρίνει η πρόβλεψη του μοντέλου με την πραγματική τιμή της κλάσης. Αν το μοντέλο προβλέψει σωστά την κλάση ενός ικανοποιητικού ποσοστού παρατηρήσεων, τότε θεωρείται επιτυχημένο και μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων.
- **Εφαρμογή (Application).** Στο στάδιο αυτό το μοντέλο αφού έχει εκπαιδευτεί και έχει κριθεί αποδεκτό, χρησιμοποιείται για τη διατύπωση προβλέψεων. Εφαρμόζεται πλέον σε νέα δεδομένα ίδιου τύπου των οποίων η κατηγοριοποίηση είναι άγνωστη και το μοντέλο καλείται να τα κατατάξει.