ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΟΙΚΟΝΟΜΙΑΣ & ΔΙΟΙΚΗΣΗΣ

Σύγκριση απόδοσης μεθόδων μέτρησης εντροπίας πληροφορίας σε πρότυπα προβλήματα απόφασης

Λάσκαρης Γεώργιος: Α.Μ.: 23110055

ΕΠΙΒΛΕΠΩΝ: Δούνιας Γεώργιος

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ:

Δούνιας Γεώργιος

Βασιλάκης Παναγιώτης

Κούτρας Βασίλειος

ΧΙΟΣ, *2021*

Δήλωση Ακαδημαϊκής Ακεραιότητας

Ο υπογράφων προπτυχιακός φοιτητής του Πανεπιστημίου Αιγαίου της Πολυτεχνικής Σχολής του τμήματος Μηχανικών Οικονομίας και Διοίκησης (Ο.ΔΙ.Μ.)

Λάσκαρης Γεώργιος του Ιωάννη (Α.Μ.: 23110055)

Δηλώνω υπεύθυνα ότι:
Έχω διαβάσει και κατανοήσει τους κανόνες για τη λογοκλοπή και τον τρόπο σωστής αναφοράς των πηγών που περιέχονται στον Οδηγό συγγραφής διπλωματικών εργασιών του ΤΜΟΔ. Δηλώνω ότι, από όσα γνωρίζω, το περιεχόμενο της παρούσας διπλωματικής εργασίας είναι προϊόν δικής μου δουλειάς και υπάρχουν αναφορές σε όλες τις πηγές που χρησιμοποίησα.

Ημερομηνία: Νοέμβριος, 2021
Λάσκαρης Γεώργιος

Ευχαριστίες

Με την ολοκλήρωση της παρούσας πτυχιακής εργασίας θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γεώργιο Δούνια για την καθοδήγηση, την υποστήριξη και τις πολύτιμες συμβουλές που μου παρείχε. Θα ήθελα, επιπλέον, να ευχαριστήσω τον κ. Βαγγέλη Καραμπότση για την πολύτιμη βοήθειά του και τις κατατοπιστικές του απαντήσεις στα ερωτήματά μου.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Περίληψη

Η συγκεκριμένη εργασία παρουσιάζει τους κυριότερους ορισμούς της εντροπίας και εστιάζει στην ανάλυση των βασικότερων μέτρων εντροπίας καθώς και των πρακτικών χρήσεών τους. Στη συνέχεια, επιλέχθηκαν κάποιοι ορισμοί και μέτρα, τα οποία συγκρίθηκαν μέσα από μια συγκεκριμένη επαναληπτική πειραματική διαδικασία, σε μια σειρά πρότυπων προβλημάτων της βιβλιογραφίας από τον χώρο της λήψης αποφάσεων. Με σκοπό την εξαγωγή συμπερασμάτων για τη χρήση του καταλληλότερου ορισμού ή μέτρου της εντροπίας ανά τύπο και χαρακτηριστικά προβλήματος, πραγματοποιήθηκαν συγκρίσεις απόδοσης μεταξύ των μεθόδων.

Οι εργασίες θα εκπονηθούν στο εργαστήριο ΔΕΛΑΠ με τη συνεπίβλεψη των καθηγητών Δούνια - Βασιλάκη που μελετούν τα συγκεκριμένα θέματα καθώς και τη συμμετοχή του κ. Κούτρα στην τριμελή επιτροπή.

Το θέμα της παρούσας διπλωματικής εργασίας αφορά τη συγκριτική ανάλυση μεθόδων μέτρησης εντροπίας πληροφορίας και αλγορίθμων ταξινόμησης σε ένα σύνολο πραγματικών δεδομένων. Συγκεκριμένα, έγινε προσπάθεια να αντιμετωπιστεί το πρόβλημα της διάγνωσης του καρκίνου του μαστού, από δεδομένα που συνέλλεξε ο Dr. William H. Wolberg στο Πανεπιστήμιο του Wisconsin. με στόχο να μπορέσει ο εξεταζόμενος αλγόριθμος ταξινόμησης να προβλέψει σωστά μία νέα παρατήρηση ως φυσιολογική ή μη. Μελετήθηκαν διάφορες μέθοδοι επιλογής χαρακτηριστικών προκειμένου να μειωθεί η πολυπλοκότητα και ο τυχόν θόρυβος στα δεδομένα του προβλήματος και δόθηκε έμφαση στην απόδοση-αποτελεσματικότητα των επιλεγμένων μεθόδων ταξινόμησης. Για την ανάλυση των δεδομένων και την παραγωγή των αποτελεσμάτων, χρησιμοποιήθηκε το πρόγραμμα Jypiter Notebook (anaconda3) το οποίο διαθέτει μία μεγάλη βιβλιοθήκη αλγόριθμων και παρέχει ένα εύχρηστο περιβάλλον στο χρήστη. Σε αυτή τη μελέτη, χρησιμοποιήθηκε μία βάση δεδομένων η οποία περιέχει 613 παρατηρήσεις και 10 χαρακτηριστικά τα οποία περιγράφουν την κάθε περίπτωση και έγινε σύγκριση της προβλεπτικής ικανότητας των αλγόριθμων ταξινόμησης σε ένα πρόβλημα δύο κλάσεων. Η εργασία χωρίζεται σε οκτώ (8) κεφάλαια. Στο πρώτο κεφάλαιο γίνεται μία εισαγωγή και γενική αναφορά στη θεωρία πληροφορίας. Στο δεύτερο κεφάλαιο γίνεται μία εκτενής αναφορά στο μέτρο ποσότητας πληροφορίας, την εντροπία. Αρχικά, προσδιορίζονται οι κυριότερες αρχές και ιδιότητες που τη χαρακτηρίζουν και έπειτα αναλύονται τα βασικά είδη εντροπίας προκειμένου

στα κεφάλαια που ακολουθούν να γίνει κατανοητή η χρησιμότητά τους σε βασικά προβλήματα της εποχής μας. Στο τρίτο κεφάλαιο γίνεται αναφορά στα προβλήματα απόφασης και η εφαρμογή που έχει η εντροπία πάνω σε αυτά. Στο τέταρτο κεφάλαιο παρουσιάζεται το πρόβλημα της ταξινόμησης και πώς αυτό συνδέεται άμεσα με τη διαδικασία της επιλογής χαρακτηριστικών και γίνεται περιγραφή βασικών αλγόριθμων ταξινόμησης. Το πέμπτο κεφάλαιο περιλαμβάνει την πειραματική ανάλυση καθώς και όλα τα αποτελέσματα που προέκυψαν από το Jypiter Notebook. Στο έκτο κεφάλαιο, παρουσιάζονται τα αποτελέσματα των μεθόδων εντροπίας mutual information, Shannon Entropy και Renyi Entropy που εφαρμόστηκαν πάνω στο dataset Wisconsin breast cancer. Στο έβδομο κεφάλαιο παρουσιάζονται τα αποτελέσματα των αλγορίθμων ταξινόμησης στο αρχικό dataset καθώς επίσης και τα υποσύνολα που δημιουργήθηκαν έπειτα από την εφαρμογή συγκεκριμένων μεθόδων εντροπίας. Στο τελευταίο κεφάλαιο γίνεται αναφορά στον τρόπο εργασίας και τη μεθοδολογία που ακολουθήθηκε καθ’ όλη τη διάρκεια της έρευνας και παρατίθενται ορισμένα βασικά συμπεράσματα.

Εφαρμόστηκαν διάφοροι συνδυασμοί μεθόδων επιλογής χαρακτηριστικών και αλγόριθμων ταξινόμησης με καλά αποτελέσματα όπως προέκυψε μέσα από την έρευνα. Παρατηρήσαμε ότι η επιλογή χαρακτηριστικών δεν βελτίωσε την ακρίβεια των ταξινομητών. Εφαρμόζοντας πολλαπλή επικύρωση ten fold cross validation και μελετώντας το μέτρο Recall παρατηρείται πως η επιλογή χαρακτηριστικών, με τη μέθοδο Shannon Entropy, βελτίωσε την απόδοση και των δύο ταξινομητών (95.8%) ενώ η μέθοδος Renyi Entropy κράτησε την απόδοση τους στα ίδια επίπεδα (95.4%).

Εισαγωγή

Είναι γενικά αποδεκτό, περισσότερο σήμερα παρά ποτέ, ότι ζούμε στην εποχή που όλες οι δραστηριότητές μας έχουν να κάνουν με τον ένα ή τον άλλο τρόπο με τους ηλεκτρονικούς υπολογιστές και κατ' επέκταση με τον κόσμο της πληροφορίας. Τα άτομα σαν οντότητες και σαν επιχειρηματικοί φορείς, αλλά και όλοι ανεξαιρέτως οι κλάδοι επιστημών αντλούν αλλά και παρέχουν πληροφορίες καθημερινά και ασταμάτητα, κινούμενοι σε ένα παγκόσμιο δίκτυο τεράστιου όγκου δεδομένων.

Η εξόρυξη δεδομένων είναι μια από τις αναπτυσσόμενες επιστήμες στον κόσμο. Οι αλγόριθμοι εξόρυξης δεδομένων με βάση τις λειτουργίες τους μπορούν να χωριστούν σε διάφορες κατηγορίες όπως η ταξινόμηση, η επιλογή χαρακτηριστικών και η ομαδοποίηση. Μία από τις πιο σημαντικές λειτουργίες είναι η επιλογή χαρακτηριστικών, η οποία αναπτύσσεται όλο και περισσότερο και πολλοί ερευνητές παρέχουν ποικιλία αλγορίθμων για την αντιμετώπιση αυτής της συνάρτησης τα τελευταία χρόνια. Οι αλγόριθμοι επιλογής χαρακτηριστικών χρησιμοποιούνται κυρίως για την απόκτηση πιο ακριβών και ισχυρών αλγορίθμων μηχανικής μάθησης μαζί με τη μείωση του χρόνου υπολογισμού. Μια άλλη αναπτυσσόμενη επιστήμη είναι οι τεχνικές λήψης αποφάσεων πολλαπλών κριτηρίων που έχει επίσης ποικιλία μεθόδων. Σε αυτή την εργασία, χρησιμοποιούμε την μέθοδο της Εντροπίας, η οποία έχει ως λειτουργία τη στάθμιση των κριτηρίων για την επιλογή των κατάλληλων χαρακτηριστικών.

1 Η Θεωρία της Πληροφορίας

1.1 Βασικές Έννοιες

Η Θεωρία της Πληροφορίας είναι ο κλάδος των εφαρμοσμένων μαθηματικών που έχει ως αντικείμενο την ποσοτικοποίηση της πληροφορίας. Εμπνευστής της συγκεκριμένης θεωρίας ήταν ο Claude Shannon που προσδιόρισε τα θεμελιώδη όρια της επεξεργασίας σήματος σε εφαρμογές όπως η συμπίεση, η αποθήκευση και η μεταφορά δεδομένων.[1] Η διεύρυνση της συγκεκριμένης θεωρίας, από τη στιγμή της θεμελίωσής της μέχρι και σήμερα, είναι πάρα πολύ μεγάλη και βρίσκει εφαρμογή σε πολλούς τομείς, όπως στην επαγωγική στατιστική, σε δίκτυα εκτός των δικτύων επικοινωνίας, στην επεξεργασία της φυσικής γλώσσας, στη νευροβιολογία στην εξέλιξη και τη λειτουργία μοριακών κωδικών στην οικολογία, στη φυσική, στους κβαντικούς υπολογιστές, στην ανίχνευση λογοκλοπής καθώς και σε άλλους χώρους ανάλυσης δεδομένων.

Ένα από τα βασικά μέτρα της πληροφορίας είναι η εντροπία τής, που εκφράζεται συνήθως μετρώντας το μέσο αριθμό των bits που χρειάζονται για να αποθηκευτεί ή να μεταβιβαστεί ένα σύμβολο σε ένα μήνυμα.[2] Γενικά, η εντροπία της πληροφορίας ποσοτικοποιεί την αβεβαιότητα που εμπλέκεται στην πρόβλεψη μιας τιμής μιας τυχαίας μεταβλητής.

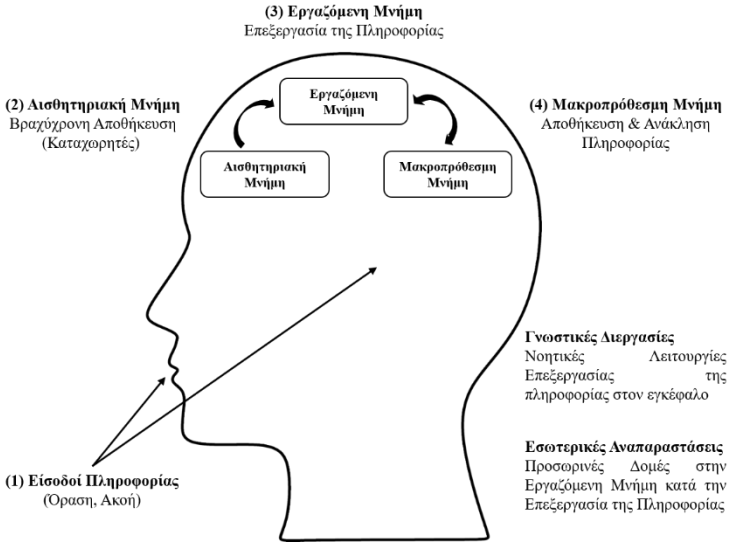
Τα κυριότερα θέματα, που σχετίζονται με τη Θεωρία της Πληροφορίας, είναι α) η ποσότητα συντακτικής πληροφορίας (εντροπία), β) οι μονάδες μέτρησης της εντροπίας, γ) η ροή πληροφορίας σε κανάλια και δ) τα όρια της ποσότητας πληροφορίας που είναι δυνατόν να μεταδοθούν, δηλαδή η χωρητικότητα των καναλιών ή ρυθμός μετάδοσης της πληροφορίας. Άλλα θέματα είναι η κατασκευή συστημάτων επεξεργασίας της πληροφορίας, τα συστήματα επικοινωνίας πληροφορίας κ.ο.κ..

Δύο βασικά θεωρήματα της Θεωρίας της Πληροφορίας, είναι α) το θεώρημα του Shannon για τον πηγαίο κώδικα, το οποίο καθιερώνει ότι, κατά μέσο όρο, ο αριθμός των δυαδικών ψηφίων που χρειάζονται για να αναπαρασταθεί το αποτέλεσμα ενός αβέβαιου γεγονότος δίνεται από την εντροπία της πληροφορίας του και β) το θεώρημα του Shannon, για τα θορυβώδη κανάλια μεταφοράς, το οποίο αναφέρει πως η αξιόπιστη επικοινωνία είναι

[1] (Shannon 1948)
[2] (Zenil 2020)

δυνατή σε θορυβώδη κανάλια με την προϋπόθεση ότι ο συντελεστής της επικοινωνίας είναι κάτω από ένα ορισμένο όριο.[3]

Η Θεωρία της Πληροφορίας είναι ένας ευρύς επιστημονικός κλάδος, με εξίσου ευρείες και «βαθιές» εφαρμογές, όπως προαναφέρθηκε. Εντούτοις, ο τομέας, όπου η Θεωρία της Πληροφορίας βρίσκει την πιο πλατιά αποδοχή και είναι κάτι περισσότερο από απαραίτητη, είναι αυτός της Θεωρίας της Κωδικοποίησης. Τέλος, μερικοί άλλοι (πιο σύγχρονοι) τομείς, στους οποίους χρησιμοποιείται η Θεωρία της Πληροφορίας είναι η ανάκτηση πληροφοριών, η συλλογή πληροφοριών καθώς και η σύνθεση μουσικής.



Εικόνα 1.1. Σχηματική αναπαράσταση της θεωρίας επεξεργασίας πληροφορίας

Η εντροπία μιας τυχαίας μεταβλητής X με συνάρτηση πιθανότητας p(X) ορίζεται από την εξίσωση [1.1]

$$\mathrm{H(x)} = -\sum_{\mathrm{x}} \mathrm{p(x)} \log_2 \mathrm{p(x)} \qquad [1.1]$$

[3] (Gallager 1968)

1.2 **Ιστορική Αναδρομή**

Η έννοια της πληροφορίας, πριν τα μέσα του 20^{ου} αιώνα, αποτελούσε (κατά βάση) ποιοτική και αφηρημένη έννοια. Ωστόσο, η ραγδαία ανάπτυξη της μετάδοσης πληροφορίας (information transmission) οδήγησε στην αντίληψη και μελέτη της πληροφορίας υπό ένα νέο πρίσμα, αυτό της ποσοτικής και μαθηματικής της υπόστασης.

Πρώτος ο Samuel Morse το 1837 ανέπτυξε το αλφάβητο Morse για τη μετάδοση μηνυμάτων με βέλτιστο τρόπο σε μεγάλη απόσταση. Έπειτα, ο Graham Bell το 1903 ανακάλυψε τον πολλαπλό τηλέγραφο και τέλος ο Thomas Edison το 1874 εισήγαγε την τετραπλή κωδικοποίηση (quadruplex) για να αυξήσει τον ρυθμό μετάδοσης της πληροφορίας. Τα συγκεκριμένα επιτεύγματα ώθησαν την επιστημονική κοινότητα στη μελέτη βέλτιστων τρόπων αναμετάδοσης της πληροφορίας και συνετέλεσαν στη θεμελίωση ενός νέου τομέα της Θεωρίας Επικοινωνίας, αυτού της Θεωρίας Πληροφορίας.

Η δημοσίευση του Harry Nyquist το 1924, με τίτλο «*Certain Factors Affecting Telegraph Speed*», περιέχει ένα θεωρητικό τμήμα που ποσοτικοποιεί τη νοημοσύνη και την ταχύτητα γραμμής κατά την οποία μπορεί να μεταδοθεί από ένα σύστημα επικοινωνίας, δίνοντας την εξίσωση [1.2]

$$W = K \log m \qquad [1.2]$$

όπου,

W: η ταχύτητα μετάδοσης της νοημοσύνης

m: ο αριθμός από διαφορετικά επίπεδα τάσης για να επιλέξουμε σε κάθε χρονικό βήμα

K: σταθερά.

Η δημοσίευση του Ralph Hartley το 1928 με τίτλο «*Transmission of Information*», χρησιμοποιεί τη λέξη *πληροφορία* ως μια μετρήσιμη ποσότητα που αντανακλά την ικανότητα του δέκτη να ξεχωρίζει μια σειρά από σύμβολα από μια οποιαδήποτε άλλη σειρά συμβόλων, έτσι ποσοτικοποιεί την πληροφορία σύμφωνα με την εξίσωση [1.3].

$$H = \log S^n = n \log S \qquad [1.3]$$

όπου,

S: ο αριθμός από πιθανά σύμβολα

n: ο αριθμός από σύμβολα σε μια μετάδοση.

Η φυσική μονάδα πληροφοριών ήταν επομένως το δεκαδικό ψηφίο, όπου αργότερα μετονομάστηκε σε Hartley (προς τιμήν του ερευνητή) ως μονάδα ή κλίμακα ή μέτρηση της πληροφορίας.

Ο Alan Turing το 1940 χρησιμοποίησε παρόμοιες ιδέες ως κομμάτι μιας στατιστικής ανάλυσης για το «σπάσιμο» των αλγορίθμων κρυπτογράφησης των Γερμανών κατά το Δεύτερο Παγκόσμιο πόλεμο.

Πολλά από τα μαθηματικά, που σχετίζονται με τη Θεωρία της Πληροφορίας, αναπτύχθηκαν εντός του πεδίου της θερμοδυναμικής από τον Ludwig Boltzmann και Josiah Willard Gibbs. Οι σχέσεις μεταξύ πληροφορίας & θεωρητικής εντροπίας, συμπεριλαμβανομένων των σημαντικών συνεισφορών του Rolf Landauer τη δεκαετία του 1960, διερευνώνται στο άρθρο «*Entropy in thermodynamics and information theory*».

Το ορόσημο γεγονός που καθιέρωσε τον επιστημονικό κλάδο της Θεωρίας της Πληροφορίας και τον έφερε σε άμεση παγκόσμια προσοχή ήταν η δημοσίευση του Claude Shannon, με τίτλο «*A mathematical Theory of Communication*» στο επιστημονικό περιοδικό «*Bell System Technical Journal*» τον Ιούλιο και τον Οκτώβριο του 1948.[4] Πριν από αυτή τη δημοσίευση περιορισμένες πληροφορίες (θεωρητικές ιδέες) είχαν εξελιχθεί στα εργαστήρια Bell labs που ήταν όλες υποθετικά σιωπηρά γεγονότα ίσης πιθανότητας. Στην επαναστατική και πρωτοποριακή αυτή εργασία, η οποία είχε ολοκληρωθεί στα Bell Labs στα τέλη του 1944, ο Claude Shannon για πρώτη φορά εισήγαγε το ποσοτικό και ποιοτικό μοντέλο επικοινωνίας ως μια στατιστική διαδικασία όπου στηρίζεται η θεωρία της πληροφορίας, εισάγοντας τον εξής ισχυρισμό: «*το θεμελιώδες πρόβλημα της επικοινωνίας είναι αυτό της αναπαραγωγής σε ένα σημείο, είτε ακριβώς είτε κατά προσέγγιση, ενός μηνύματος που επιλέχθηκε σε ένα άλλο σημείο*» και στηριζόμενοι στον συγκεκριμένο ισχυρισμό κατέληξαν στην επινόηση:

- της εντροπίας πληροφοριών και της Redundancy (ο αριθμός των δυαδικών ψηφίων που χρησιμοποιούνται για τη μετάδοση ενός μηνύματος μείον τον αριθμό των δεκαδικών ψηφίων της πραγματικής πληροφορίας σε ένα μήνυμα) μιας πηγής, και τη σημασία της μέσω του θεωρήματος του πηγαίου κώδικα,

[4] (Stone 2015).

- της κοινής πληροφορίας και της χωρητικότητας ενός θορυβώδους καναλιού, συμπεριλαμβανομένης και της υπόσχεσης για τέλεια (χωρίς απώλειες) επικοινωνία που δίνεται από το θεώρημα του θορύβου καναλιού,
- του πρακτικού αποτελέσματος του νόμου Shannon-Hartley για τη χωρητικότητα του καναλιού ενός Gaussian καναλιού, καθώς επίσης,
- του bit, ένας νέος τρόπος οπτικής για την πιο θεμελιώδη μονάδα πληροφοριών.

Το μέτρο πληροφορίας όπως ορίστηκε από τον Claude Shannon, δίνεται από τη σχέση:

$$H(X) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i \qquad [1.4]$$

Τέλος, ο Shannon (1948) διαπίστωσε ότι όσο λιγότερο θόρυβο έχει ένα κανάλι μεταφοράς πληροφορίας, τόση περισσότερη πληροφορία θα μεταδώσει. Αντιστρόφως, όσο αυξάνεται ο θόρυβος (αταξία), τόσο λιγότερη πληροφορία μεταδίδεται. Συνεπώς, η πληροφορία είναι μέτρο της εσωτερικής τάξης ενός συστήματος και αντιστρόφως ανάλογη της αταξίας. Δηλαδή, αφού η εντροπία αποτελεί μέτρο αταξίας ενός συστήματος, είναι αντιστρόφως ανάλογη της πληροφορίας.

1.3 **Μορφές Πληροφορίας & Διαλειτουργικότητα**

Η πραγματική διάσταση της πληροφορίας εξαρτάται, κυρίως, από το δεδομένο γενικό πλαίσιο. Πριν γίνει μία περεταίρω ανάλυση των τύπων πληροφορίας, θα πρέπει να γίνει μία αναφορά στις κυριότερες έννοιες που σχετίζονται με αυτή και μία από αυτές είναι η διαλειτουργικότητα.

Η **Διαλειτουργικότητα** χαρακτηρίζει την ικανότητα συνεργασίας ενός συστήματος με άλλα συστήματα, χωρίς να καταβληθεί ιδιαίτερη προσπάθεια. Ένα σημαντικό χαρακτηριστικό της διαλειτουργικότητας είναι η ετερογένεια, που είναι μία άλλη πλευρά της εντροπίας. Μία άλλη έννοια, που είναι σχετική με το αντικείμενο της πληροφορίας, είναι αυτή των μεταδεδομένων, δηλαδή τα δεδομένα με τα οποία περιγράφονται άλλα δεδομένα που αποτελούν μια πηγή πληροφορίας. Ουσιαστικά, δίδεται η δυνατότητα, η μορφή αυτής της ετερογένειας, να δεικτοδοτηθεί. Σε ένα επίπεδο διαμορφωμένης πληροφορίας, το οποίο

αποτελεί το τεκμήριο, η ετερογένεια πληροφορίας μπορεί να διακριθεί στη συντακτική και τη σημασιολογική.

Η συντακτική πληροφορία έχει σχέση με τα σύμβολα και τις σχέσεις μεταξύ τους, τα οποία σύμβολα αποτελούν τα μηνύματα.

Σε αυτό το είδος πληροφορίας, λόγω της αυξημένης εκφραστικότητας της ανθρώπινης γλώσσας, τα ίδια δεδομένα μπορεί να παριστάνονται με διαφορετικό τρόπο, κάθε φορά. Η αυξημένη αυτή εκφραστικότητα, οδηγεί σε μία μεγάλη γκάμα κωδικοποιήσεων δεδομένων π.χ. διαφορετικά λειτουργικά υπολογιστικά συστήματα, διαφορετικές κλίμακες, διαφορετικές πλατφόρμες κ.α.[5]

Η σημασιολογική πληροφορία σχετίζεται με τη σημασιολογική ετερογένεια κατά Goh, όπου παρατηρούνται ζητήματα όπως σύγκρουσης αντιθέσεων και κλιμάκωσης αντιθέσεων, διαφορά ονομάτων κ.λπ.. Στη σύγκρουση αντιθέσεων τα αντικείμενα της πληροφορίας φαίνεται, σε πρώτη φάση, πως έχουν την ίδια έννοια αλλά διαφέρουν ουσιαστικά μεταξύ τους. Στην κλιμάκωση αντιθέσεων γίνεται μία επεξήγηση των διαφορών, μέσω του καθορισμού μιας αξίας τοποθετημένης στη βάση διαφορετικών συστημάτων αναφοράς. (π.χ. αξία νομίσματος). Τέλος, στην περίπτωση διαφοράς των ονομάτων έχουμε τη χρησιμοποίηση διαφορετικών σχημάτων στα ονόματα πληροφοριών (π.χ. συνώνυμα/ομώνυμα)[6].

Τη λύση στα προβλήματα που δημιουργούνταν από τις διάφορες μορφές ετερογένειας έδωσε η γλώσσα σημειοθέτησης γενικής χρήσης, XML (Extensible Markup Language - Επεκτάσιμη γλώσσα σημειοθέτησης). Η συγκεκριμένη γλώσσα θεωρείται επεκτάσιμη γλώσσα, γιατί δίνει τη δυνατότητα στους χρήστες να καθορίσουν οι ίδιοι τις ετικέτες τους. Ο κύριος στόχος της είναι να βοηθήσει στο διαμοιρασμό δομημένων δεδομένων μέσω του διαδικτύου. Εκτός αυτού, η XML χρησιμοποιείται για την κωδικοποίηση εγγράφων αλλά και για τη μετατροπή των δεδομένων σε δυαδική μορφή ώστε για να μπορούν να αποθηκευτούν τοπικά ή και να μεταφερθούν μέσω ενός δικτύου. Η XML σχεδιάσθηκε, αρχικά, ως υποσύνολο της γλώσσας SGML (Standard Generalized Markup Language), με σκοπό να είναι σχετικά ευανάγνωστη από το χρήστη. Αργότερα, όμως, η δυνατότητα για καθοδηγούμενη περιγραφή του εγγράφου μέσω του DTD (Document Type Definition) αν και οδήγησε σε διαλειτουργικές λύσεις, αύξησε την πολυπλοκότητα σε θέματα εξειδικευμένης περιγραφής

[5] IGNOU (2017)
[6] IGNOU (2017)

εγγράφου όπως για παράδειγμα στην ιατρική πληροφορία, όπου το XML δένδρο θεωρείται ότι είναι πολύ μεγάλο.[7] Με αφορμή το συγκεκριμένο ζήτημα, δημιουργήθηκε μία άλλη μετρική εντροπία που αποσκοπεί στον καθορισμό τριών μεταβλητών: α) τη πυκνότητα των πληροφοριών, β) το μέγεθος του εγγράφου, και γ) τη δομή της κανονικότητας. Χαμηλή πυκνότητα πληροφοριών σε ένα έγγραφο σημαίνει ότι περιέχει κυρίως διαρθρωτικές πληροφορίες, ενώ υψηλή τιμή σημαίνει ότι το έγγραφο περιέχει κυρίως πραγματικό περιεχόμενο.

Η Επαγωγική Στατιστική (ΕΣ), μεταξύ των άλλων, έχει ως αντικείμενο μελέτης και τη δημιουργία εικονικής μνήμης. Η επαγωγική στατιστική είναι μία σειρά από μεθόδους που χρησιμοποιούνται για την εξαγωγή συμπερασμάτων αναφορικά με τον πληθυσμό και τα χαρακτηριστικά του, τα οποία βασίζονται σε δεδομένα ενός δείγματος. Η Περιγραφική Στατιστική (ΠΣ) σε αντίθεση με την Επαγωγική Στατιστική, κατατάσσει τα στατιστικά δείγματα σε ομάδες και στη συνέχεια τα αναλύει σε πίνακες και διαγράμματα ανάλογα με ορισμένες χαρακτηριστικές τιμές ή συγκεκριμένες ιδιότητες. Η ανάλυση αυτή γίνεται με δύο μεθόδους, τη μονοπαραμετρική μέθοδο και την πολυπαραμετρική.[8]

Η Μονοπαραμετρική Ανάλυση περιγράφει την κατανομή μιας μόνο μεταβλητής, και συμπεριλαμβάνει τη διασπορά και την κεντρική τάση. Το σχήμα που έχει η κατανομή μπορεί επίσης να περιγραφεί με διάφορους δείκτες, όπως είναι η κύρτωση ή η ασυμμετρία. Άλλα χαρακτηριστικά που περιλαμβάνει η κατανομή μιας μεταβλητής είναι η γραφική απεικόνιση ή η απεικόνιση με μορφή πίνακα, ανάλογα με τις ιδιότητες των δειγμάτων.

Η Πολυπαραμετρική Ανάλυση χρησιμοποιείται στην περίπτωση που ένα δείγμα αποτελείται από περισσότερες της μίας μεταβλητές. Η πολυπαραμετρική ανάλυση χρησιμοποιείται για να περιγράψει τη σχέση μεταξύ δύο η περισσότερων μεταβλητών. Οι πιο πάνω διαδικασίες καθορίζουν τεχνικές ομαδοποίησης των δειγμάτων προκειμένου να ορίσουν νέες προσεγγίσεις ανάλυσης των δεδομένων που έχουν σαν βάση την αρχή της ομαδοποίησης.

Ανάλυση Ομάδων (Clusters) ή Ομαδοποίηση (Clustering) ονομάζεται η κατάταξη ενός συνόλου δεδομένων, με τέτοιον τρόπο ώστε τα δεδομένα της ίδιας ομάδας να μοιάζουν μεταξύ τους όσο το δυνατόν περισσότερο (πρέπει, τουλάχιστον, να τα συνδέει μια κοινή

[7] (Papaleo 2014)
[8] (Χάλκος 2011)

ιδιότητα), από ότι τα δεδομένα άλλων ομάδων (clusters).[9] Η μεθοδολογία ομαδοποίησης είναι ένα από τα βασικότερα εργαλεία του κλάδου της Εξόρυξης Δεδομένων και εφαρμόζεται σε πολλούς επιστημονικούς τομείς της Μηχανικής Μάθησης, όπως η Ανάλυσης Εικόνας, η Αναγνώρισης Μορφών (patterns, motives), η Ανάκτηση Πληροφοριών κ.λπ..

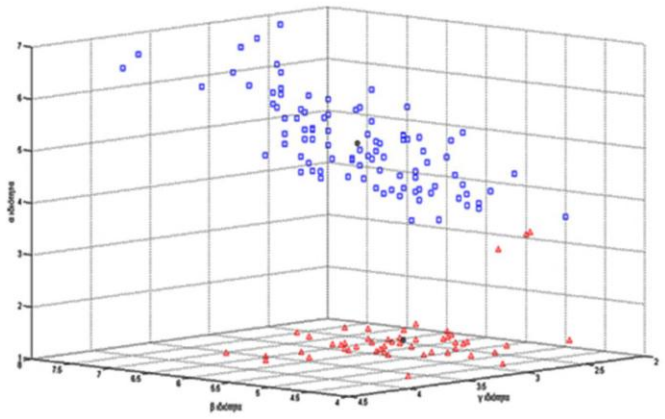
Η Ανάλυση Ομάδων πραγματοποιείται με πολλές διαφορετικές αλγοριθμικές τεχνικές. Εφαρμόζονται διάφοροι αλγόριθμοι, που διαφέρουν μεταξύ τόσο ως προς την κατανόηση του περιεχόμενου της ομαδοποίησης, όσο και προς την διαδικασία του εντοπισμού και στη συνέχεια του αποτελεσματικού διαχωρισμού των δεδομένων. Αναφορικά, υψηλής ομογενοποίησης ομάδες είναι αυτές που χαρακτηρίζονται από μικρές «αποστάσεις» μεταξύ των μελών τους, αυτές που έχουν περιοχές με υψηλής πυκνότητας του χώρου των δεδομένων κλπ.

Η Ομαδοποίηση Δεδομένων είναι ένα συνηθισμένο πρόβλημα πολυπαραμετρικής βελτιστοποίησης. Οι αλγόριθμοι ομαδοποίησης, οι τιμές απόστασης, τα κατώφλια πυκνότητας, ο αριθμός συμπλεγμάτων που αναμένονται και οι λοιπές παράμετροι, εξαρτώνται από το σύνολο των δεδομένων και του σκοπού χρήσης των αποτελεσμάτων.

Εκτός αυτών, η Ανάλυση Ομάδων, δεν είναι μία «αυτόματη εργασία», αλλά μια διαδικασία «εξόρυξης γνώσης» που επαναλαμβάνεται. Πολλές φορές θεωρείται αναγκαία η τροποποίηση της προκαταρκτικής επεξεργασίας και κάποιων άλλων ή ακόμα και όλων των ιδιοτήτων και των παραμέτρων, προκειμένου να επιτευχθεί το προσδοκώμενο αποτέλεσμα (Εικόνα 1.2).

[9] Aldenderfer et al 1984)

Εικόνα 1.2. Ομαδοποίηση Δεδομένων Πολυπαραμε/κών Χαρ/κών με Κανόνα τη Συνοχή



Πηγή: Πούλος (2015), σ. 5.

Ομαδοποιήσεις που πιθανόν να βρέθηκαν με διαφορετικούς αλγόριθμους, ενδέχεται να διαφέρουν σημαντικά όσον αφορά τις ιδιότητές τους. Τα παρακάτω μοντέλα ομαδοποίησης μας βοηθούν στην κατανόηση των ανομοιοτήτων που υπάρχουν, μεταξύ των διαφόρων αλγορίθμων.[10]

- Τα Κεντροειδή Μοντέλα: Σε αυτούς τους αλγορίθμους κάθε ομάδα αναπαρίσταται με ένα διάνυσμα.
- Τα Μοντέλα Συνδεσιμότητας: Είναι μοντέλα Ιεραρχικής Ομαδοποίησης που στηρίζονται σε απόσταση συνδεσιμότητας.
- Μοντέλα Πυκνότητας: Τα μοντέλα DBSCAN και OPTICS βασίζονται πάνω σε συμπλέγματα σύνδεσης περιοχών με υψηλή πυκνότητα στο χώρο των δεδομένων.
- Μοντέλα Κατανομής: Δίνεται η δυνατότητα να μοντελοποιηθούν χρησιμοποιώντας στατιστικές κατανομές.
- Μοντέλα Υποχώρων: Μοντελοποιούνται και με τα δύο μέλη του συμπλέγματος και με πρόσθετες ίσως κοινές ιδιότητες.

Δεν μπορούμε να πούμε πως υπάρχει συγκεκριμένα κάποιος αντικειμενικά «σωστός» αλγόριθμος ομαδοποίησης, αλλά τις περισσότερες φορές είναι ανάγκη κάποιος να επιλέγεται

[10] (Omran 2007)

πειραματικά, εκτός και αν υπάρχει ένας ιδιαίτερος λόγος που καθιστά απαραίτητα τη χρήση ενός συγκεκριμένου μοντέλου. Κάποιος αλγόριθμος που έχει σχεδιαστεί για ένα είδος μοντέλου, δεν έχει πολλές πιθανότητες επιτυχίας σε ένα γενικό σύνολο δεδομένων βασιζόμενο σε διαφορετικού είδους μοντέλα.

1.4 **Πηγή πληροφορίας**

Ως πηγή πληροφορίας ονομάζεται κάθε σύστημα που στην έξοδο του παράγει πληροφορία. Η πληροφορία εξάγεται με τη μορφή συμβόλων. Τα διακεκριμένα αυτά σύμβολα που αναπαριστάνουν την πληροφορία μίας πηγής, ονομάζονται Αλφάβητο της Πηγής. Το αλφάβητο της πηγής συμβολίζεται ως:

$$X = \{x_1, x_2, \dots, x_N\} \quad [1.5]$$

Το πλήθος N των συμβόλων μπορεί να είναι είτε πεπερασμένο είτε άπειρο.

Κάθε διατεταγμένη ακολουθία συμβόλων ονομάζεται λέξη και κάθε διατεταγμένη ακολουθία λέξεων, ονομάζεται μήνυμα. Κάθε σύμβολο έχει μια πιθανότητα να εμφανιστεί στην έξοδο της πηγής. Η πηγή έχει συνολικά μια κατανομή πιθανοτήτων του συνόλου των συμβόλων που έχει στη διάθεσή της[11].

$$PX = \{px_1, px_2, \dots, px_N\} \quad [1.6]$$

Η πηγή συμβολίζεται από η δυάδα που σχηματίζεται από το αλφάβητό και την κατανομή πιθανοτήτων των συμβόλων (X,PX). Τα διαδοχικά σύμβολα που εκπέμπει μία πηγή είναι ανεξάρτητα μεταξύ τους. Το σύμβολο που εκπέμπει μία πηγή σε μία οποιαδήποτε χρονική στιγμή είναι ανεξάρτητο από αυτό που είχε εκπέμψει σε προηγούμενες επιλογές (διακριτή πηγή χωρίς μνήμη).

1.4.1 Ποσότητα Πληροφορίας Πηγής

Τα μηνύματα που παράγουν οι πηγές πληροφορίας αποτελούνται από ακολουθίες συμβόλων. Οι αποστολείς και οι τελικοί παραλήπτες ενδιαφέρονται κυρίως για τα μηνύματα, ενώ τα επικοινωνιακά συστήματα ασχολούνται κατά κύριο λόγο με τα σύμβολα που

[11] (Lombardi 2016)

απαρτίζουν τα αντίστοιχα μηνύματα. Δηλαδή, τα επικοινωνιακά συστήματα ενδιαφέρονται για την ποσότητα της πληροφορίας των συμβόλων που παράγονται από την πηγή..

1.4.2 Επεξεργασία της Πληροφορίας

Η επεξεργασία της πληροφορίας έχει ασαφή χαρακτηριστικά, τα οποία προέρχονται κυρίως από τη διάχυση της πληροφορίας, με την ευρεία έννοια. Στην ευρεία διάχυση πραγματοποιείται η χωρίς φορμαλισμούς επεξεργασία της πληροφορίας που μας οδηγεί στο πρόβλημα που λέγεται διαλειτουργικότητα. Το πρόβλημα αυτό απαιτεί λύση που έχει να κάνει με την ενοποίηση (Integration) και την προτυποποίηση (Standardization) των πληροφοριών. Πώς, όμως, μπορεί να αναλυθεί η σημασιολογική επεξεργασία της πληροφορίας αν δεν αναλυθούν πρώτα τα ζητήματα ερμηνείας της διάχυσής της στο επίπεδο εντροπίας; Τη λύση στις ανωτέρω μορφές ετερογένειας, δίνει η γλώσσα σημειοθέτησης γενικής χρήσης (Extensible Markup Language - Επεκτάσιμη)[12].

Η πυκνότητα πληροφορίας είναι ένας σημαντικός παράγοντας που καθορίζει το είδος του εγγράφου. Για παράδειγμα, μία χαμηλή πυκνότητα πληροφοριών σε ένα έγγραφο, μας δείχνει ότι στο συγκεκριμένο έγγραφο περιέχονται κυρίως διαρθρωτικές πληροφορίες, ενώ σε μία υψηλή τιμή παίρνουμε την πληροφορία πως το έγγραφο έχει κυρίως πραγματικό περιεχόμενο.

Οι τεχνικές ταξινόμησης και οι μεθοδολογίες ψηφιακών πληροφοριών στον τομέα του πολιτισμού (υλικό από αρχεία, μουσεία και βιβλιοθήκες) αποτελούν ένα σημαντικό αντικείμενο έρευνας της Επιστήμης της Πληροφορίας. Οι σημερινές μεθοδολογίες έχουν να κάνουν κυρίως με τεχνικές παραγωγής μεταδεδόμενων. Πρωταρχικός στόχος εδώ είναι η μονοσήμαντη δομική περιγραφή εγγράφων, η οποία βελτιώνει την έκφραση, την διάθεση, την μεταφορά και την αποθήκευση πληροφοριών. Ταυτόχρονα, η εφαρμογή XML σαν γλώσσα έκφρασης πληροφοριών έρχεται να λύσει προβλήματα ετερογένειας στην κωδικοποίηση δεδομένων. Η σύγχρονη πρακτική σχετικά με την οργάνωση και τη δημιουργία κανόνων συσχέτισης μεταξύ της γλώσσας XML και άλλων γλωσσών οντολογιών, (π.χ. RDF και OWL), έχει σαν στόχο τη σημασιολογική περιγραφή του περιεχομένου των δομημένων εγγράφων και κατέχει την πρώτη θέση στην παραγωγή σημασιολογικών δικτύων, με την πρακτική επίλυσης συντακτικής ετερογένειας.[13]

[12] (Πούλος 2015)
[13] (Πούλος 2015)

1.4.3 Μεταφορά της Πληροφορίας & Επικοινωνιακό Μοντέλο

Σε κάθε επικοινωνιακή διαδικασία, πραγματοποιείται μία ροή πληροφορίας μεταξύ αποστολέα και αποδέκτη. Η πληροφορία αυτή παίρνει διάφορες μορφές, π.χ. μουσική, λέξεις, ηλεκτρισμό, εικόνες κ.λπ.. Η μεταφορά της πληροφορίας, στη γενική περίπτωση, επιτυγχάνεται μέσω ενός δικτύου μετάδοσης.

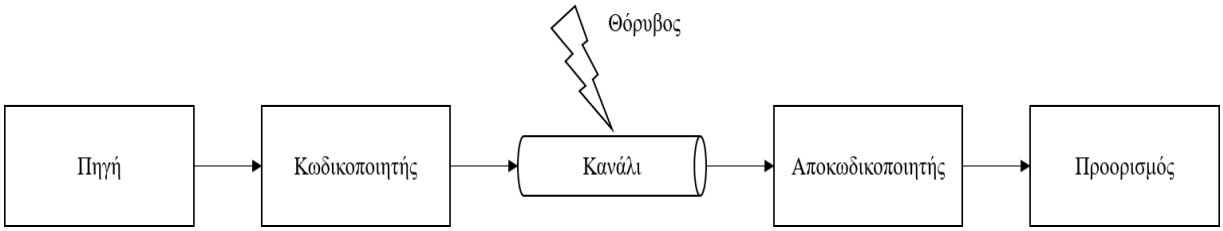
Τα βασικά μέρη ενός επικοινωνιακού μοντέλου είναι η πηγή πληροφορίας, ο αποστολέας, το δίκτυο ή αλλιώς κανάλι μετάδοσης και φυσικά ο προορισμός δηλαδή, ο παραλήπτης.

Πολύ σημαντικό ρόλο παίζει η αποθήκευση της πληροφορίας. Η αποθήκευση μπορεί να θεωρηθεί σαν μέρος του καναλιού ή του δικτύου μετάδοσης, αν και ουσιαστικά δεν είναι ακριβώς ζήτημα μετάδοσης. Η πληροφορία κατά τη διαδικασία της μετάδοσής της μπορεί να αλλοιωθεί εξαιτίας του θορύβου πάνω στο κανάλι.[14]

Η μεταφορά της πληροφορίας δε θα πρέπει να έχει σφάλματα. Αυτό μπορούμε να το πετύχουμε με τη διόρθωση των σφαλμάτων που θα παρατηρήσουμε. Μια τέλεια μεταφορά χωρίς σφάλματα είναι κατά κανόνα αδύνατη όταν έχει να κάνει με σήματα πολυμέσων. Αυτό που μπορεί να γίνει σε αυτές τις περιπτώσεις, είναι να τεθούν απαιτήσεις σχετικά με το αποδεκτό μέγεθος της απόκλισης μεταξύ του σήματος που λαμβάνει ο αποδέκτης και αυτού που έχει αποστείλει ο μεταδότης. Ανάλογα με τις απαιτήσεις ως προς την ποιότητα μεταφοράς της πληροφορίας επιλέγεται το κατάλληλο κανάλι/μέσο μεταφοράς και επιβάλλονται οριακές συνθήκες προσαρμογής σε αυτό. Οι σχεδιαστές των επικοινωνιακών συστημάτων, σήμερα, επιδιώκουν κυρίως την ελαχιστοποίηση των απωλειών της πληροφορίας στο κανάλι και την όσο το δυνατόν καλύτερη ποιότητα επανάκτησης πληροφορίας η οποία έχει προσβληθεί από θόρυβο. Για το λόγο αυτό χρησιμοποιούνται διάφορες τεχνικές κωδικοποίησης για τον αποστολέα και αντίστοιχα άλλες τεχνικές αποκωδικοποίησης για την πλευρά του αποδέκτη. Λαμβάνοντας υπόψη την κωδικοποίηση και την αποκωδικοποίηση, μπορούμε να καταλήξουμε στη γενική δομή ενός επικοινωνιακού μοντέλου (Εικόνα 1.3) και ενός λεπτομερούς επικοινωνιακού μοντέλου (Εικόνα 1.4 &1.5).

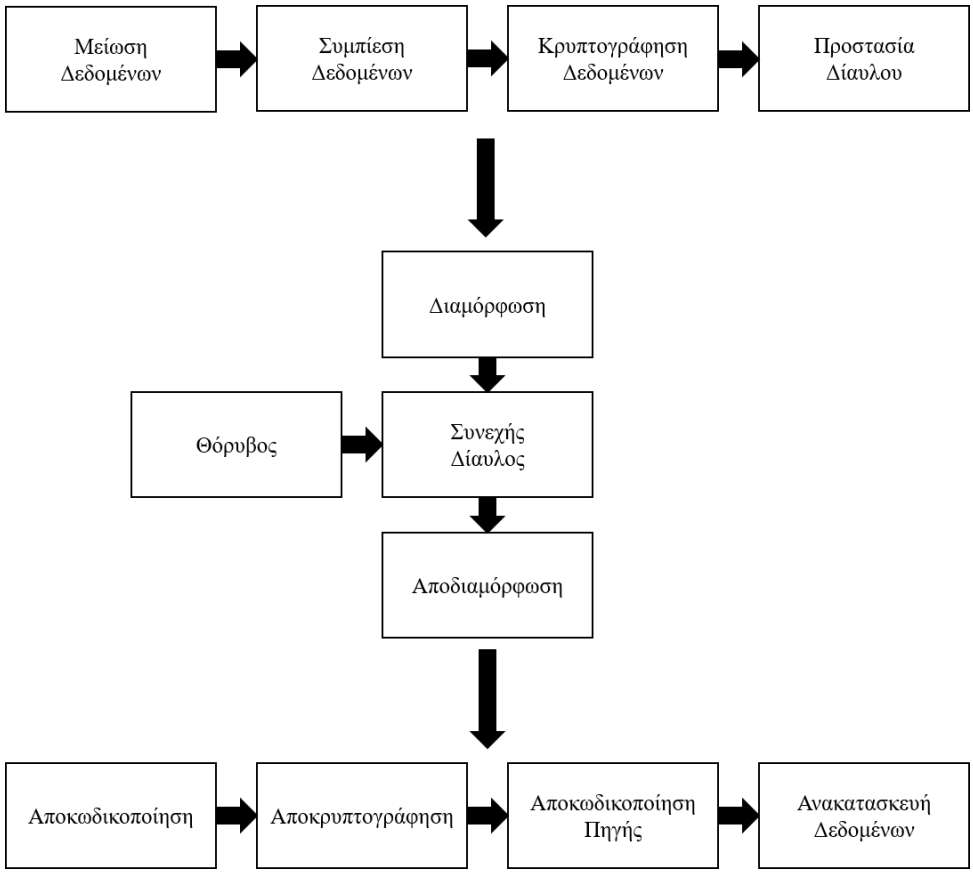
[14] (Shannon & Weaver 1949)

Εικόνα 1.3. Βασικό Επικοινωνιακό Διάγραμμα

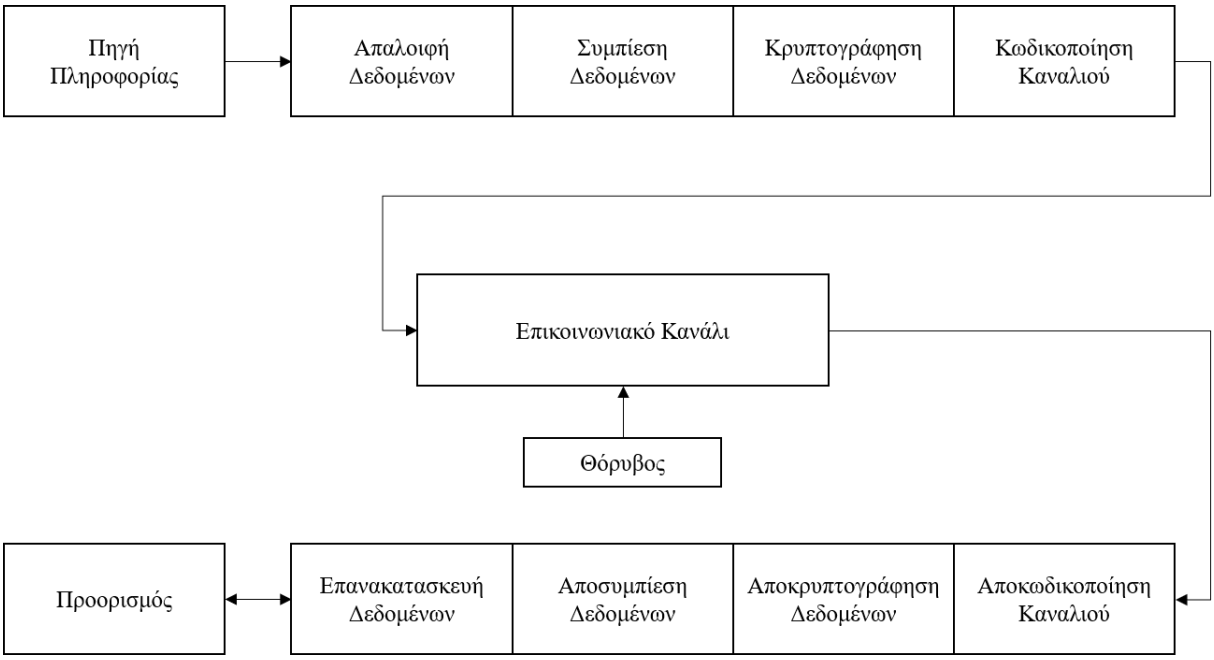


Πηγή: Shannon & Weaver (1949), p. 7.

Εικόνα 1.4. Σχεδιάγραμμα λεπτομερούς επικοινωνιακού μοντέλου (1)



Εικόνα 1.5. Σχεδιάγραμμα λεπτομερούς επικοινωνιακού μοντέλου (2)



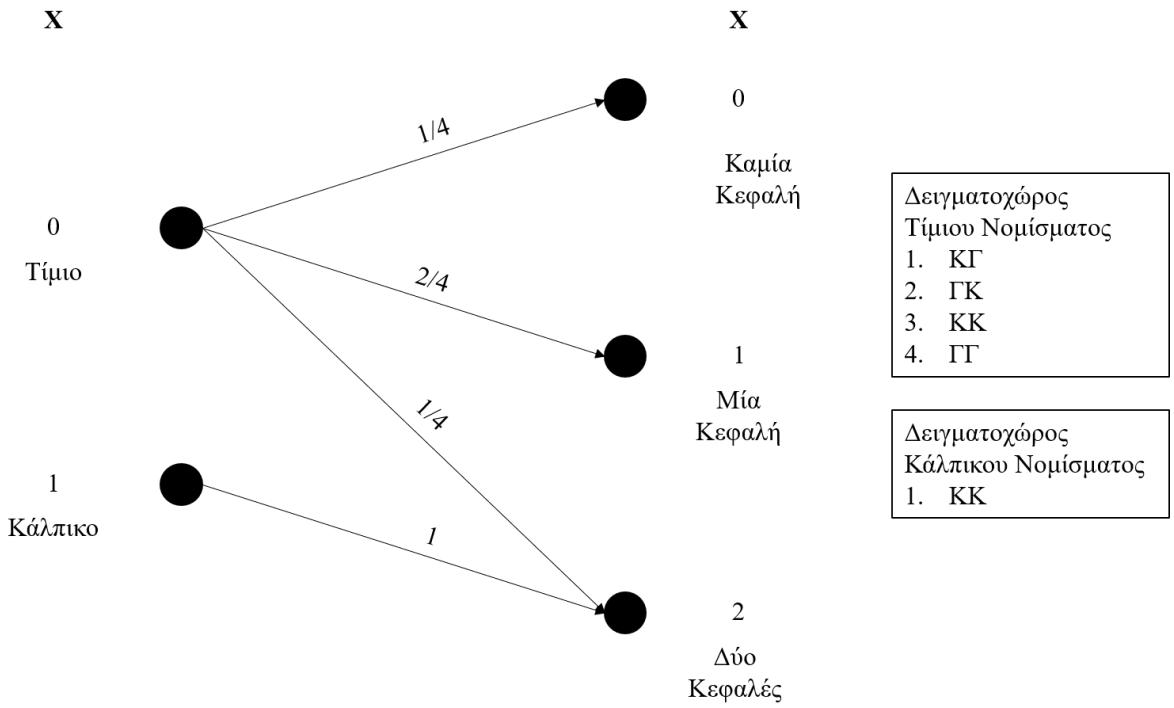
1.5 **Διαπληροφορία**

Η πληροφορία που αποκομίζουμε για την πηγή X γνωρίζοντας το αποτέλεσμα μίας άλλης πηγής Y ονομάζεται διαπληροφορία και μας δείχνει το κατά πόσο μειώθηκε η αβεβαιότητα για την έξοδο της πηγής X, όταν γνωρίζουμε την έξοδο της πηγής Y. Η διαπληροφορία ουσιαστικά φανερώνει το ποσό (bits/symbol) της πληροφορίας που μεταφέρθηκε από την είσοδο ενός διαύλου στην έξοδο του[15]. Στην είσοδο και στην έξοδο στην περίπτωση της διαπληροφορίας λειτουργούν δυο πηγές πληροφορίας. Στη διαπληροφορία μειώνεται η αβεβαιότητα από την πλευρά του παρατηρητή.

Μπορούμε να διαπιστώσουμε τα παραπάνω με ένα παράδειγμα (Εικόνα 1.6) δύο νομισμάτων. Το ένα απ' αυτά είναι γνήσιο και το άλλο κάλπικο. Το κάλπικο έχει δύο κεφαλές. Αρχικά, ρίχνουμε το ένα νόμισμα δύο φορές και σημειώνουμε τον αριθμό των κεφαλών που προέκυψαν. Στη συνέχεια, απαντάμε στην ερώτηση για το πόσες κεφαλές παρατηρήσαμε στο τέλος του πειράματος. Από τον αριθμό των κεφαλών μπορεί να έχουμε

[15] (Learned-Miller 2013)

την απάντηση σχετικά με το ποιο νόμισμα επιλέχθηκε τελικά. Εάν ο αριθμός των κεφαλών είναι μικρότερος από 2, τότε το νόμισμα που επιλέχθηκε ήταν το γνήσιο (τίμιο) νόμισμα. Εάν, όμως, ο αριθμός κεφαλών ήταν ίσος με 2, τότε είναι πιθανόν να επιλέχθηκε το κάλπικο νόμισμα.



Εικόνα 1.6. Παράδειγμα Διαπληροφορίας

2 Εντροπία της Πληροφορίας

2.1 Μέτρα Ποσότητας Πληροφορίας

Η μέση ποσότητα πληροφορίας ονομάζεται εντροπία. Κεντρικό ρόλο στη Θεωρία της Πληροφορίας παίζει η ίδια η έννοια της πληροφορίας. Σύμφωνα με τη συγκεκριμένη θεωρία, η πληροφορία έχει ποσοτικό χαρακτήρα και συνεπώς διαφέρει σημαντικά από το εννοιολογικό περιεχόμενο που της αποδίδουμε στην καθημερινή μας ζωή[16]. Το περιεχόμενο της ποσοτικοποίησης της έννοιας της πληροφορίας και ο ορισμός ενός κατάλληλου μέτρου για τον υπολογισμό τής απασχόλησε τον Hartley το 1928. Ο Hartley, κατά τη μελέτη των τηλεγραφικών επικοινωνιών, διαπίστωσε πως όσο πιο μεγάλη είναι η πιθανότητα εμφάνισης ενός γεγονότος, τόσο πιο μικρή είναι η αβεβαιότητα για το αν θα συμβεί το συγκεκριμένο γεγονός. Στην περίπτωση που το γεγονός συμβεί, η πληροφορία που θα λάβουμε θα είναι μικρή. Από τα παραπάνω είναι φανερό ότι συνηθισμένα γεγονότα, όπως για παράδειγμα «*Σήμερα ο ήλιος ανέτειλε*» συνοδεύονται από μικρή ποσότητα πληροφορίας, ενώ σπάνια γεγονότα, όπως παραδείγματος χάρη «*Σήμερα έγινε ολική έκλειψη ηλίου*» συνοδεύονται από μεγάλη ποσότητα πληροφορίας. Αυτό συμβαίνει, επειδή το δεύτερο γεγονός έχει μικρή πιθανότητα να συμβεί σε σχέση με το πρώτο, που είναι βέβαιο. Επομένως, αν X είναι ένα τυχαίο γεγονός με πιθανότητα p(X) και I(X) είναι η συνάρτηση του μέτρου της πληροφορίας του X, τότε η I(X) θα πρέπει να ικανοποιεί τις παρακάτω ιδιότητες[17]:

- Όταν η πιθανότητα να συμβεί ένα γεγονός είναι μονάδα, τότε η ποσότητα της μεταφερόμενης πληροφορίας είναι μηδενική. Αυτό σημαίνει ότι δεν χρειάζεται η διαβίβαση του μηνύματος, αφού το γεγονός είναι σίγουρο ότι θα συμβεί, δηλαδή:

$$I_x = 0, \text{όταν } P_X = 1 \quad [2.1]$$

- Η πληροφορία ενός γεγονότος είναι ένα μη αρνητικό μέγεθος, αφού ισχύει:

$$0 \leq P_x \leq 1 \quad [2.2]$$

  δηλαδή:

$$I_x \geq 0 \text{ όταν } 0 \leq P_x \leq 1 \quad [2.3]$$

- Όσο πιο απίθανο είναι να συμβεί ένα γεγονός, τόσο περισσότερη πληροφορία λαμβάνουμε από την πραγματοποίησή του, δηλαδή:

[2.4]

[16] Arndt (2003)
[17] Shannon (1948)

$$I_x \geq I_y \text{ όταν } P_x \leq P_y$$

- Αν τα γεγονότα X και Y είναι ανεξάρτητα με αντίστοιχες πιθανότητες P_x,P_y, τότε το μέτρο της πληροφορίας του γεγονότος εμφάνισης και των δύο επιμέρους γεγονότων είναι ίσο με το άθροισμα των δύο επιμέρους μέτρων πληροφορίας, δηλαδή:

$$I_{xy} = I_x + I_y \text{ όταν } P_{xy} = P_x P_y \quad [2.5]$$

Η παραπάνω σχέση αποδεικνύεται ως εξής:

$$I_{xy} = -\log_b(P_x P_y) \Rightarrow I_{xy} = -\log_b(P_x) - \log_b(P_y) \Rightarrow I_{xy} = I_x + I_y \quad [2.6]$$

Έτσι υιοθετούμε τον παρακάτω ορισμό της πληροφορίας.

Ορισμός 1: Η πληροφορία I_x, την οποία αποκτούμε από την πραγματοποίηση ενός γεγονότος X, το οποίο έχει πιθανότητα P_x, δίνεται από τον τύπο:

$$I_x = -\log_b P_x \equiv \log_b\left(\frac{1}{P_x}\right), \text{όπου } b > 1 \quad [2.7]$$

Όπως παρατηρείται στη παραπάνω σχέση, η πληροφορία είναι ένα αδιάστατο μέγεθος, ενώ η βάση b του λογαρίθμου μπορεί να επιλεγεί ελεύθερα, αρκεί $b > 1$.

Η μονάδα μέτρησης της πληροφορίας καθορίζεται ανάλογα με τη βάση υπολογισμού του λογαρίθμου. Έτσι, όταν χρησιμοποιείται ο φυσικός λογάριθμος, τότε η μονάδα είναι το nat, ενώ, όταν χρησιμοποιείται ο δεκαδικός λογάριθμος, η μονάδα είναι το Hartley ή decit. Η επικρατέστερη μονάδα μέτρησης της πληροφορίας είναι το bit[18]. Ο λόγος που επικράτησε έναντι των άλλων μονάδων είναι η χρησιμοποίηση του δυαδικού συστήματος στους υπολογιστές.

Η εντροπία του Shannon μπορεί να χρησιμοποιηθεί και για τον ορισμό άλλων μέτρων της πληροφορίας που αναδεικνύουν τις σχέσεις μεταξύ δύο τυχαίων μεταβλητών X και Y:

- Τη σχετική εντροπία (relative entropy), η οποία μετράει την ομοιότητα των X και Y.
- Τη σύνθετη εντροπία (joint entropy), η οποία μετράει τη συνολική πληροφορία των X και Y.

[18] Vajapeyam (2014)

- Την υπό συνθήκη εντροπία (conditional entropy), η οποία μετράει την πληροφορία του X, όταν η Y είναι γνωστή και αντιστρόφως.
- Την αμοιβαία-κοινή εντροπία (mutual entropy), η οποία μετράει τη σχέση των X και Y, υπό την έννοια ότι μας δείχνει πόσο μειώνεται η πληροφορία του X όταν μαθαίνουμε το Y και αντιστρόφως.
- Την υπό συνθήκη αμοιβαία πληροφορία (conditional mutual information), η οποία μετράει την αναμενόμενη αμοιβαία πληροφορία των X και Y όταν είναι γνωστή μία τρίτη μεταβλητή Z.

2.2 **Είδη Εντροπίας της Πληροφορίας**

2.2.1 Εντροπία κατά Shannon

Το μέτρο ποσότητας πληροφορίας του Hartley, δεν λαμβάνει υπόψη διαφορετικές πιθανότητες για την επιλογή των συμβόλων που απαρτίζουν ένα μήνυμα. Η εισαγωγή, από τον Shannon, της έννοιας της πιθανότητας στον ορισμό του μέτρου ποσότητας πληροφορίας έθεσε τις βάσεις για την ανάπτυξη της σύγχρονης Θεωρίας της Πληροφορίας. Ο Shannon γενίκευσε, λοιπόν, τον ορισμό της ποσότητας πληροφορίας του Hartley, επιτρέποντας διαφορετικές πιθανότητες εμφάνισης των συμβόλων σε μηνύματα και κατ' επέκταση και των διαφόρων μηνυμάτων. Η συσχέτιση της έννοιας της πιθανότητας με τον ορισμό του μέτρου ποσότητας πληροφορίας είναι εύλογη. Αν θεωρήσουμε ένα τυχαίο πείραμα με δειγματοχώρο του οποίου τα γεγονότα είναι ισοπίθανα, τότε υπάρχει μεγάλη αβεβαιότητα για το αποτέλεσμα. Αντίθετα, αν ο δειγματοχώρος έχει ένα στοιχείο με πολύ μεγάλη πιθανότητα, τότε το να συμβεί αυτό το γεγονός προσφέρει πολύ λιγότερη πληροφορία απ' ότι το να συμβεί ένα από τ' άλλα γεγονότα.

Αν X είναι μία διακριτή τυχαία μεταβλητή με δειγματοχώρο $X = \{x_1, x_2, \dots, x_n\}$ και συνάρτηση πιθανότητας μάζας $p(x_i)$, τότε η μέση ποσότητα πληροφορίας (ή μέση πληροφορία ή μέσο πληροφορικό περιεχόμενο) της X, H(X), δίνεται από τη σχέση[19]:

$$\mathrm{H(x)} = -\sum_{\mathrm{x}} \mathrm{p(x)} \log_2 \mathrm{p(x)} \qquad [2.8]$$

[19] Shannon (1948)

Η μέση πληροφορία ονομάζεται εντροπία.

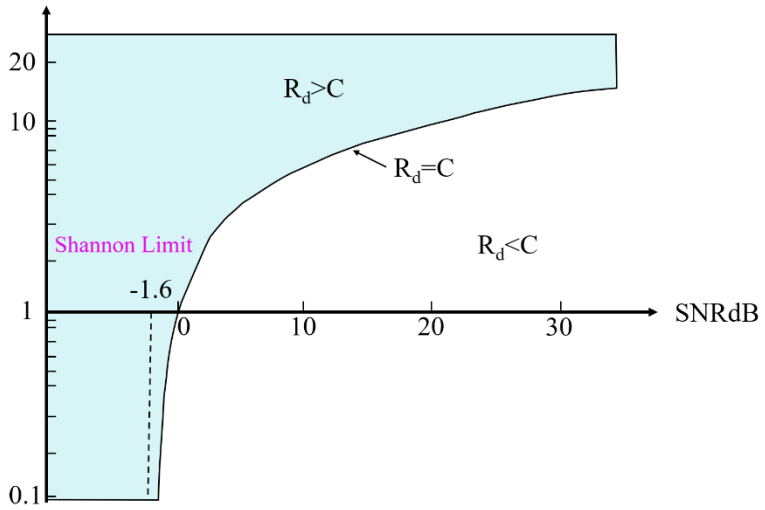
Στην περίπτωση μιας διακριτής τυχαίας μεταβλητής X με δύο ενδεχόμενα, π.χ. εκπομπή ενός από δύο δυνατά μηνύματα και πιθανότητες αυτών p και (1–p), αντίστοιχα, η εντροπία είναι:

$$H(X) = -p \log p - (1-p) \log(1-p) \qquad [2.9]$$

Όπως μπορούμε να συνάγουμε από τον ορισμό της εντροπίας, η ποσότητα πληροφορίας (ή το πληροφορικό περιεχόμενο) ενός γεγονότος χ της τυχαίας μεταβλητής X είναι ίσο με τον αρνητικό λογάριθμο της πιθανότητας εμφάνισής του p(x), δηλαδή ίσο με $(-\log p(x))$. Επομένως, η ποσότητα πληροφορίας ενός γεγονότος είναι αντιστρόφως ανάλογη της πιθανότητας εμφάνισής του.

Ο Shannon δηλώνει ότι ένα μέτρο της ποσότητας πληροφορίας H(p) που περιέχεται σε μια σειρά γεγονότων $p_1 \ldots p_n$ θα πρέπει να ικανοποιεί τρεις απαιτήσεις:

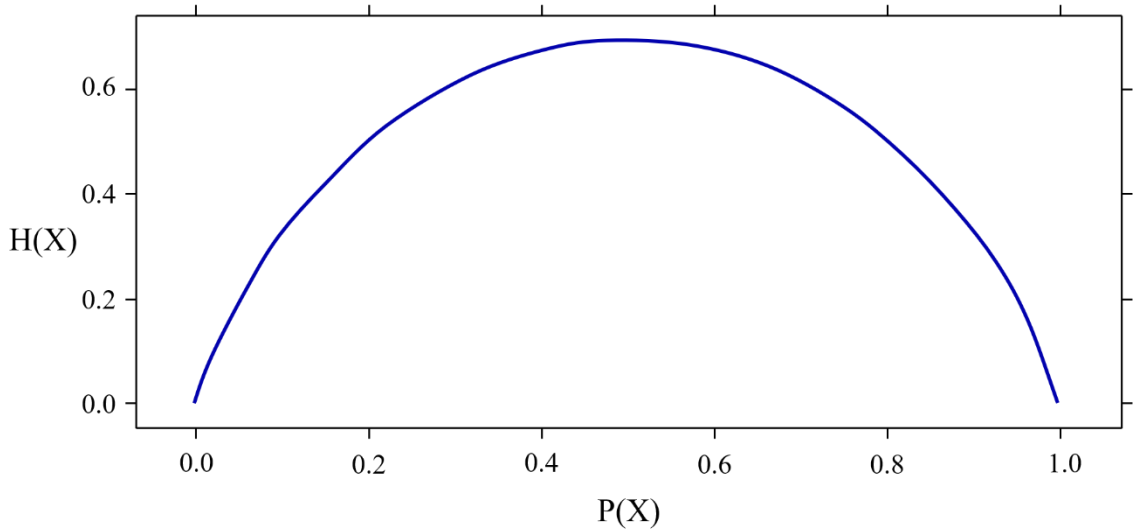
- Η πρέπει να είναι συνεχής στο p_i.
- Εάν όλα τα p_i είναι ισοπίθανα, έτσι ώστε $p_i = \frac{1}{N}$ τότε η H θα πρέπει να είναι μια μονοτονική αύξουσα συνάρτηση του N.
- Η πρέπει να είναι προσθετική.



Εικόνα 2.1. Γραφική αναπαράσταση του θεωρήματος Shannon

Η γραφική παράσταση του παρακάτω σχήματος, μας δείχνει τη συμπεριφορά της εντροπίας ως συνάρτηση της πιθανότητας p. (Η μονάδα μέτρησης της εντροπίας είναι το bit, δηλαδή ο λογάριθμος είναι με βάση το 2.)

Εικόνα 2.2. Γραφική απεικόνιση της εντροπίας μιας Bernulli κατανομής



Πηγή: Edwards (2008), p. 15

Παρατηρούμε, λοιπόν, στη γραφική παράσταση ότι η μέση πληροφορία παίρνει τη μέγιστη τιμή, που ισούται με ένα, όταν τα δύο γεγονότα μπορούν να συμβούν με την ίδια πιθανότητα. Από την άλλη πλευρά, αν p = 1 ή p = 0, τότε η εντροπία είναι 0, αφού το τελικό αποτέλεσμα (η έκβαση του πειράματος) είναι βέβαιο.

2.2.2 Σχετική Εντροπία

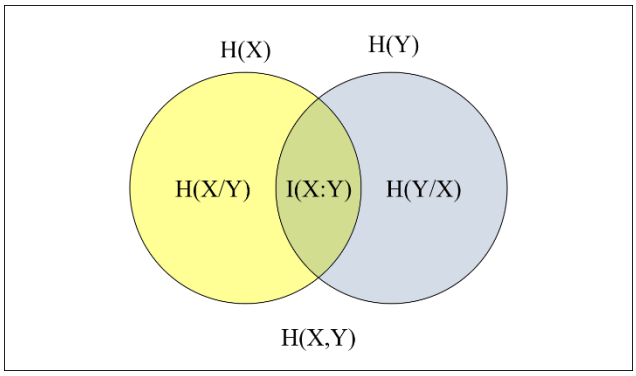
Η σχετική εντροπία ορίζεται ως εξής:

$$H\left(X\middle|Y\right) = -\sum_{x,y} p(x)\log\big(p(y)\big) - H(X) = \sum_{x,y} p(x)\log\left(\frac{p(x)}{p(y)}\right) \qquad [2.10]$$

Η εντροπία αυτή αναπαριστά τη διαφορά μεταξύ της αναμενόμενης πληροφορίας που λαμβάνεται από τα γεγονότα της Y που κατανέμονται σύμφωνα με τη X (δηλαδή

$\sum_{x,y} p(x) \log(p(y))$) και της αναμενόμενης πληροφορίας από τα γεγονότα της X (δηλαδή H(X)) και είναι πάντοτε θετική. Στη περίπτωση που X=Y, η σχετική εντροπία είναι 0. Επίσης, η σχετική εντροπία αυξάνεται όσο αυξάνεται η «απόσταση» μεταξύ των X και Y[20].

Η σχετική εντροπία είναι μια έννοια πολύ μεγάλης σημασίας για την κλασική στατιστική μηχανική του Gibbs και χρησιμοποιείται πολύ συχνά στην Κβαντική Θεωρία Πληροφορίας, διότι πολλά σημαντικά αποτελέσματα της τελευταίας βασίζονται στη μονοτονία της. Είναι, λοιπόν, η πιο κατάλληλη έκφραση της πληροφορίας, διότι η τελευταία είναι ένα καθαρά σχετικό γεγονός. Αυτό, γιατί η απροσδιοριστία σε μια μεταβλητή μετριέται πάντα ως προς μια άλλη μεταβλητή. Η εντροπία Shannon είναι μια ειδική περίπτωση της σχετικής εντροπίας. Πράγματι η εντροπία Shannon μιας τυχαίας μεταβλητής είναι η σχετική εντροπία ως προς μια κατάσταση που γνωρίζουμε με απόλυτη βεβαιότητα, δηλαδή H(X)=H(X||Y), όπου P(Y=y)=1 για κάποια τιμή του y. Στο παρακάτω σχήμα η σχετική εντροπία είναι το κίτρινο και το θαλασσί τμήμα.



Εικόνα 2.3. Γραφική Απεικόνιση της Σχετικής Εντροπίας

2.2.3 Σύνθετη Εντροπία

Η σύνθετη εντροπία δυο τυχαίων μεταβλητών, είναι απλά η εντροπία της σύνθετης κατανομής τους[21].

Σχέση μεταξύ εντροπίας και αμοιβαίας πληροφορίας:

$$P\{EF\} = (P\{E_k F_k\}) \rightarrow H(X,Y) = \sum_{k=1}^{n} \sum_{j=1}^{m} p_{kj} \log p_{kj} \quad [2.11]$$

$$[2.12]$$

$$[2.13]$$

[20] Cover and Thomas (1991)
[21] Cover and Thomas (1991)

$$P\{E\} = (P\{E_k\}) \rightarrow H(X) = \sum_{k=1}^{n} p\{x_k\} \log p\{x_k\}$$

$$P\{F\} = \left(P\{F_j\}\right) \rightarrow H(Y) = \sum_{j=1}^{m} p\{y_k\} \log p\{y_k\}$$

2.2.4 Υπό Συνθήκη Εντροπία

Η εντροπία αυτή μετράει την πληροφορία που λαμβάνουμε από την γνώση των Y όταν το X είναι ήδη γνωστό[22]:

$$H(Y|X) \equiv \sum_{x \in X} p(x) H(Y|X = x) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x)$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)} \Rightarrow$$

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x)}{p(x,y)} \qquad [2.14]$$

όπου $p(x|y) = p(x,y)/p(x)$

Στην περίπτωση που η τυχαία μεταβλητή Y, καθορίζεται τελείως μέσω της X, ισχύει ότι $H(Y|X) = 0$. Εάν, πάλι, τα X και Y είναι πλήρως ανεξάρτητα το ένα από το άλλο, τότε $H(Y|X) = H(Y)$, αφού η πληροφορία που αντλούμε από το Y δεν μειώνεται λόγω της γνώσης του X. Μπορούμε, τέλος, να περιγράψουμε τη σύνθετη κατανομή των X και Y περιγράφοντας πρώτα το X και μετά το Y με δεδομένο το X. Αυτή η κίνηση μας οδηγεί στη διατύπωση μιας σχέσης μεταξύ των συνθέτων και των υπό συνθήκη εντροπιών. Μέσω του κανόνα της αλυσίδας έχουμε ότι:

$$H(x,y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x) p(y|x)$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x)$$

$$= -\sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \Rightarrow$$

[2.15]

[22] Ζορκάδης (2002)

$$H(x,y) = H(X) + H(Y|X)$$

Ισοδύναμα ισχύει ότι:

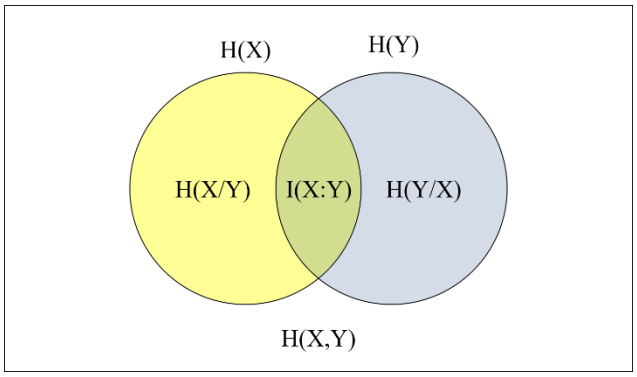
$$\log p(X,Y) = \log p(X) + \log p(Y|X) \quad [2.16]$$

2.2.5 Αμοιβαία-Κοινή Εντροπία

Η κοινή εντροπία μεταξύ δύο μεταβλητών X και Y είναι η διαφορά της πληροφορίας που λαμβάνουμε από το X από αυτή που παίρνουμε από το X όταν είναι γνωστό το Y[23]:

$$I(X:Y) = H(X) - H(X|Y)$$

Βλέποντας το Σχήμα 2-4 παρακάτω γίνεται πιο κατανοητό. Η κοινή εντροπία είναι ο μηνίσκος επικάλυψης των δύο κύκλων. Αν δεν υπάρχει συσχέτιση των X και Y , τότε αυτή είναι προφανώς 0. Αν υπάρχει πλήρης, τότε είναι ίση με την πληροφορία του X ή του Y (αφού είναι ίσες).



Σχήμα 2-4 Γραφική απεικόνιση της αμοιβαίας πληροφορίας

[23] Ζορκάδης (2002)

Σημειωτέων, ότι η κοινή εντροπία είναι συμμετρική ως προς τις δύο μεταβλητές, δηλαδή ισχύει:

$$\mathrm{I}(\mathrm{Y}:\mathrm{X}) = \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{Y}|\mathrm{X}) = \mathrm{I}(\mathrm{X}:\mathrm{Y}) \qquad [2.17]$$

Παρατηρούμε, λοιπόν, ότι η κοινή εντροπία είναι η διαφορά στον αριθμό των bits που χρειαζόμαστε για να εκφράσουμε τα X και Y ξεχωριστά απ' ότι σαν σύνθετη κατανομή, δηλαδή:

$$\mathrm{I}(\mathrm{X}:\mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{X},\mathrm{Y}) \qquad [2.18]$$

Η εντροπία της σύνθετης κατανομής της p(x,y) ορίζεται ως:

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) = -\sum_{\mathrm{x}\in\mathrm{X}}\sum_{\mathrm{y}\in\mathrm{Y}} \mathrm{p}(\mathrm{x},\mathrm{y}) \log_2[\mathrm{p}(\mathrm{x},\mathrm{y})] \qquad [2.19]$$

Όταν οι κατανομές X,Y είναι ανεξάρτητες μεταξύ τους, τότε ισχύει ότι:

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) \qquad [2.20]$$

Ενώ όταν οι μεταβλητές είναι πάνω από δυο, δηλαδή είναι $X_1, \dots, X_n$, η κοινή εντροπία ισούται με:

$$\mathrm{H}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}) = -\sum_{\mathrm{X}_1} \dots \sum_{\mathrm{X}_\mathrm{n}} \mathrm{p}(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}) \log_2[\mathrm{p}(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n})] \qquad [2.21]$$

Πίνακας 1. Μέτρα Ποσότητας Πληροφορίας και Σχέσεις μεταξύ τους

Μέτρα Ποσότητας Πληροφορίας και Σχέσεις μεταξύ τους		
Μέτρα Ποσότητας Πληροφορίας	**Μαθηματικός Ορισμός (X, Y: Τυχαίες Μεταβλητές)**	**Σχέσεις μεταξύ των διαφόρων Μέτρων**
Μέση Πληροφορία ή Εντροπία	$H(X) = -\sum_{x} p(x) \log_2 p(x)$	$H(X) \leq \log n$ $H(X) \geq \log n$
Συνδυασμένη Πληροφορίας	$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \log p(x,y)$	$H(X,y) = H(X) + H(X/Y)$ $H(X,y) = H(Y) + H(X/Y)$
Υπό Συνθήκη Πληροφορία	$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \log \frac{p(x)}{p(x,y)}$	$H(X/Y) \leq H(X)$
Αμοιβαία Πληροφορία	$I(X:Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$	$I(X:Y) = H(X) + H(X) + H(X,Y)$ $I(X:Y) = H(X) - X(X/Y)$ $I(X:Y) = H(Y) - X(Y/X)$

Πηγή: Ζορκάδης (2002, pp. 40)

2.3 Άλλα είδη εντροπίας

2.3.1 Η εντροπία του Tsallis

Η εντροπία Tsalli είναι σημαντική σε διάφορες επιστήμες και αποτελεί μία πιο εκτεταμένη μορφή από την εντροπία του Shannon. Γενικεύει την πληροφορία μέσω της γραμμικής διαδικασίας του μέσου όρου.

Ο τύπος αυτής της μεθόδου είναι ο εξής[24]:

$$S_q = k\frac{1 - \sum_{i=1}^{w} p_i^q}{q - 1} \qquad [2.22]$$

όπου,

q: ένας θετικός πραγματικός αριθμός

k: μία σταθερά την οποία θεωρούμε ίση με 1

w: είναι ο συνολικός αριθμός μικροσκοπικών διαμορφώσεων

$\{p_i\}$: το σύνολο των αντίστοιχων πιθανοτήτων το οποίο δίνεται από τον τύπο:

$$\sum_{i=1}^{n} p_i = 1 \qquad [2.23]$$

Διαπιστώνουμε ότι εάν το q→1, καταλήγουμε στην εντροπία του Shannon, γιατί:

$$\lim_{q\to 1} S_q = k \lim_{q\to 1} \frac{1 - \sum_{i=1}^{w} p_i e^{[(q-1)\ln p_i]}}{q - 1} = -k\sum_{i=1}^{w} p_i \ln p_i \qquad [2.24]$$

Συμπεραίνουμε ότι η εντροπία του Tsalli, πιθανότητας p ενός διακριτού συνόλου w έχει ως εξής:

$$S_q = \begin{cases} k\frac{1 - \sum_{i=1}^{w} p_i^q}{q - 1}, \text{αν } q \neq 1 \\ -k\sum_{i=1}^{w} p_i \ln p_i, \text{ αν } q = 1 \end{cases} \qquad [2.25]$$

[24] Tsallis (1988)

όπου S_q για το διάστημα 0<pi<1 είναι συνεχής συνάρτηση των $\{p_i\}$.

Σε ένα δεδομένο σύνολο w ισοπίθανων ενδεχομένων, για παράδειγμα pi = 1/w, το S_q είναι μία μονότονη αύξουσα συνάρτηση των w που ορίζεται ως:

$$S_q = (w^{1-q} - 1)/(1 - q) \quad [2.26]$$

Σε ανεξάρτητα συστήματα A και B η εντροπία του συστήματος A+B ικανοποιεί την αθροιστική σχέση:

$$\mathrm{S_q(A + B) = S_q(A) + S_q(B) + (1 - q)S_q(A)S_q(B)} \quad [2.27]$$

Εάν :

$$\mathrm{W = W_L + W_M,\ p_L = \sum_{i=1}^{W_L} p_i,\ \ p_M = \sum_{i=W_{L+1}}^{W} p_i,\ (\text{για αυτό } p_L + p_M = 1)} \quad [2.28]$$

έχουμε:

$$\mathrm{S_q(\{p_i\}) = S_q(p_L, p_M) + p_L^q S_q\left(\left\{\frac{p_i}{p_L}\right\}\right) + p_M^q S_q\left(\left\{\frac{p_i}{p_M}\right\}\right)} \quad [2.29]$$

2.3.2 Η Εντροπία του Rényi

Η αξιολόγηση του κινδύνου αναφορικά με μία ουδέτερη συνάρτηση πυκνότητας πιθανότητας χρήσιμης για μία μελλοντική τιμολόγηση των περιουσιακών στοιχείων (assets) μπορεί να πραγματοποιηθεί με τη μεγιστοποίηση της εντροπίας του Rényi. Εκεί που η κλασική μέθοδος με τη χρήση του μέτρου λογαριθμικής εντροπίας δεν πετυχαίνει να δώσει μία κατανομή περιγραφής τιμών των options, η εφαρμογή της μεγιστοποίησης της εντροπίας του Rényi με μερικούς περιορισμούς είναι πολύ αποτελεσματική.

Προκειμένου να εφαρμόσουμε τη μεγιστοποίηση της εντροπίας του Renyi, ορίζουμε την p(x) σαν άγνωστη συνάρτηση πυκνότητας πιθανότητας, πάνω στη θετική πραγματική γραμμή. Ακόμα, υποθέτουμε ότι η χρονική στιγμή k της κατανομής p(x) μας είναι γνωστή και την παίρνουμε από τη U. Προσπαθούμε να βρούμε εκείνη τη συνάρτηση πυκνότητας p(x), η οποία θα μεγιστοποιήσει την εντροπία του Rényi. Οι αντίστοιχοι περιορισμοί είναι $\int_0^\infty p(x)dx = 1$ και η στιγμιαία κατάσταση είναι:

$$\int_0^\infty x^k \, p(x)dx = U \qquad [2.30]$$

Με αυτούς τους περιορισμούς, λοιπόν, μεγιστοποιούμε την εντροπία του Rényi, με τον παρακάτω τύπο[25]:

$$\mathrm{H_a(p)} = \frac{1}{1-\mathrm{a}} \ln \int_0^\infty \mathrm{p^a(x)dx} \qquad [2.31]$$

με α>0 και α ≠ 1 να είναι μία σταθερά.

Για αυτή τη μεγιστοποίηση χρησιμοποιούμε τους πολλαπλασιαστές Lagrange.

Η εντροπία του Rényi ανήκει στην ομάδα των μέτρων πληροφορίας και διαπιστώνουμε υπό προϋποθέσεις, ότι μπορούμε να καταλήξουμε στην εντροπία του Shannon[26]. Για παράδειγμα, χαλαρώνοντας την τρίτη από τις απαιτήσεις του Shannon, εκείνη της προσθετικότητας, ο Renyi μπόρεσε να επεκτείνει την εντροπία του Shannon σε μια συνεχή οικογένεια μέτρων εντροπίας που υπακούν στο:

$$\mathrm{H(p)} = \frac{1}{1-\mathrm{q}} \ln \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{p_i^q} \qquad [2.32]$$

Η εντροπία του Renyi τείνει στην εντροπία του Shannon καθώς το $q \to 1$.

2.3.3 Η Kullback-Cross Εντροπία

Προκειμένου να λυθεί το πρόβλημα της διάκρισης, οι ερευνητές εισήγαγαν το μέτρο απόκλισης ή απόστασης, μεταξύ πληθυσμών, στα πλαίσια του μέτρου της πληροφορίας. Δύο πληθυσμοί, για έναν ερευνητή, διαφέρουν λιγότερο ή περισσότερο, ανάλογα με το πόσο δύσκολα μπορείς να τους διαχωρίσεις με την πιο αποτελεσματική μέθοδο.

Η Kullback-Cross εντροπία ή directed divergence ή discrimination information, δίνεται από τον τύπο[27]:

$$\mathrm{I(1:2)} = \mathrm{I_{1:2}(X)} = \int \mathrm{d\mu_1(x)} \log \frac{\mathrm{f_1(x)}}{\mathrm{f_2(x)}} = \int \mathrm{f_1(x)} \log \frac{\mathrm{f_1(x)}}{\mathrm{f_2(x)}} \mathrm{d\lambda(x)} \qquad [2.33]$$

[25] Rényi (1961)
[26] Maszczyk και Włodzisław (2008)
[27] Kullback και Leibler (1951)

Η απόσταση δύο κατανομών πιθανοτήτων 1 και 2 δεν είναι συμμετρική, δηλαδή $I_{1:2} \neq I_{2:1}$. Για να γίνει μετρική αυτή η απόσταση μπορούμε να προσθέσουμε τη $I_{1:2}(X)$ και τη $I_{2:1}(X) = \int f_2(x) \log f_2(x) / f_1(x) d\lambda(x)$ δηλαδή έχουμε :

$$J_{12}(E) = I_{1:2}(E) + I_{2:1}(E) == \frac{1}{\mu_1(E)} \int d\mu_1(x) \log \frac{f_1(x)}{f_2(x)} + \frac{1}{\mu_2(E)} \int d\mu_2(x) \log \frac{f_2(x)}{f_1(x)} =$$

$$= \int \left(\frac{f_1(x)}{\mu_1(E)} - \frac{f_2(x)}{\mu_2(E)}\right) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) \qquad [2.34]$$

Στην περίπτωση, όμως, μίας διακριτής τυχαίας μεταβλητής, ο τύπος της Kullback-Cross εντροπίας είναι:

$$I(P: Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \quad με \quad \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = \uparrow \qquad [2.35]$$

2.3.4 Η Tsallis Relative Εντροπία

Το πρόβλημα των συνεχών δοκιμών, ποικίλει στις διάφορες επιστήμες. Τα μέτρα εντροπίας είναι ένα πεδίο με βαρύνουσα σημασία στις παραμετρικές δοκιμές. Ο ερευνητής Robinson χρησιμοποίησε την εντροπία Kullback-Leibler, η οποία είναι βασισμένη πάνω στην εντροπία του Shannon για μία σύγκριση μεταξύ ανεξάρτητων και εξαρτημένων χρονοσειρών σχετικά με τις συναλλαγματικές ισοτιμίες πολλών νομισμάτων έναντι του δολαρίου.

Πριν αρκετά χρόνια, προτάθηκε, μία γενίκευση του θεωρήματος των Boltzmann-Gibbs-Shannon που απευθύνονταν σε μη εκτεταμένα συστήματα και βασίζονταν στην εντροπία.

Η Tsallis relative entropy είναι μία γενίκευση των παραπάνω θεωριών και δίνεται από τον τύπο[28]:

$$I_Q(p, p_0) = \int p(x) \frac{\left[\frac{p(x)}{p_0(x)}\right]^{q-1} - 1}{q - 1} dx = -\int p(x) \frac{\left[\frac{p_0(x)}{p(x)}\right]^{1-q} - 1}{1 - q} dx \qquad [2.36]$$

Όταν το q→1, παρατηρούμε ότι προκύπτει η Kullback Cross-entropy.

[28] Prehl και λοιποί (2012)

2.3.5 Η Fuzzy Εντροπία

Η Fuzzy εντροπία αποτελεί σημαντικό αντικείμενο της θεωρίας των συνόλων Fuzzy. Τα σύνολα Fuzzy μας γίνονται γνωστά το 1965 σε μία προσπάθεια παραγωγής μίας φόρμουλας για την αντιμετώπιση ποικίλων προβλημάτων. Στη φόρμουλα αυτή σημαντικό ρόλο έπαιξε η αοριστία η οποία προκύπτει κυρίως από ένα είδος ασάφειας παρά από κάποια στατιστική διαφοροποίηση.

Οι Luca και Termini ασχολήθηκαν πρώτοι με την ανάλυση της θεωρίας συνόλων Fuzzy για να κατανοηθεί περισσότερο η σχέση της με την κλασική θεωρία συνόλων. Σύμφωνα με τους παραπάνω ερευνητές η εντροπία αυτή μπορεί να εκληφθεί ως μέτρο ποσότητας πληροφοριών που δε σχετίζεται απαραίτητα με πειράματα τυχαία και την εισήγαγαν χωρίς να χρησιμοποιήσουν έννοιες πιθανότητας.

Στη θεωρία αυτή, οι Luca και Termini[29] κατέληξαν στο ότι η εντροπία είναι ένα μέτρο ασάφειας το οποίο δείχνει τη δυσκολία του να πάρει κάποιος μία απόφαση, εάν ένα στοιχείο ανήκει ή όχι στο σύνολο. Ένα μέτρο H(A) ενός συνόλου Fuzzy A έχει τις παρακάτω τέσσερις ιδιότητες:

- Το H(A) ελάχιστο, εάν $\mu_A(x) = 0$ ή 1 για όλα τα x.
- Το H(A) μέγιστο, εάν και μόνο αν $\mu_A(x) = 0{,}5$ για όλα τα x.
- $H(A) \geq H(A_1)$, όπου ισχύει $\mu_{A_1}(X) \leq \mu_A(x)$ αν $\mu_A(x) \leq 0{,}5$ και $\mu_{A_1}(x) \geq \mu_A(x)$ αν $\mu_A(x) \geq 0{,}5$
- H(A) = H($\bar{A}$), με $\bar{A}$ να είναι το συμπληρωματικό σύνολο του συνόλου A.

Ο τελικός τύπος είναι:

$$H_{DTE} = -k \sum_{i=1}^{n} \mu_i \log p_i + (1 - \mu_i) \log(1 - \mu_i) \qquad [2.37]$$

Οι ερευνητές Bhandari και R.Pal[30] μίλησαν για ένα νέο μέτρο πληροφορίας για να ξεχωρίσουν δύο σύνολα Fuzzy. Υπό προϋποθέσεις καταλήγουμε στη μη πιθανοτική εντροπία

[29] Luca και Termini (1972)
[30] Bhandari και R.Pal (1993)

η οποία αναπτύχθηκε από τους Luca και Termini. Χρησιμοποίησαν, επίσης, στην έρευνά τους και το μέτρο απόστασης ανάμεσα σε δύο σύνολα, το οποίο μέτρο απόστασης συνοδεύεται από ιδιότητες που προσδιορίζουν ένα μέτρο ασάφειας. Ο ορισμός της μη πιθανοτικής εντροπίας του συνόλου Fuzzy μπορεί να ληφθεί και ως επέκταση της μη πιθανοτικής εντροπίας του Rényi. Ακόμα, η εντροπία των Luca και Termini αποτελεί μία υποπερίπτωση της εντροπίας του Rényi, εάν α→1.

Ο Bart Kosko[31] είναι ένας ακόμα ερευνητής που ασχολήθηκε με την εντροπία Fuzzy. Κατά τον Bart Kosko η εντροπία είναι η αβεβαιότητα που συνδέεται με την πληροφορία και με κανένα τρόπο δε συνδέεται με τη θεωρία των πιθανοτήτων. Αναπτύχθηκε, έτσι, ένα νέο γενικό μέτρο εντροπίας που βασίζεται σε μία διαισθητική αναλογία αποστάσεων των συνόλων Fuzzy A και A^{near} και στην απόσταση των A και $A^{far} = (A^{near})^c$ ως:

$$H_{KOE}(q, A) = d^q(A, A^{near})/d^q(A, A^{far}) \qquad [2.38]$$

Δύο άλλοι ερευνητές[32], ο Nikhil R. Pal και ο James C. Bezdec αναφέρθηκαν στην Resolutional Uncertainty (RU), τη Probabilistic Uncertainty (PU), και τη Fuzzy Uncertainty (FU). Συνδυάζοντας τα παραπάνω, διατύπωσαν την άποψη ότι σε ένα πολύπλοκο σύστημα πιθανότατα εμπεριέχονται και τρία παραπάνω είδη αβεβαιότητας. Συσχέτισαν, ακόμα, διαφορετικά μέτρα προκειμένου να βοηθηθούν τους χρήστες να επιλέξουν το κατάλληλο μέτρο ανάλογα με την εφαρμογή που τους ενδιαφέρει.

Όλοι οι παραπάνω ερευνητές συνέβαλαν στην ερμηνεία της εντροπίας και της αβεβαιότητας που τη χαρακτηρίζει, η οποία αβεβαιότητα θεωρούν ότι είναι αποτέλεσμα όχι της ανεπάρκειας πληροφοριών αλλά της αοριστίας. Για τους Li και Liu[33], ειδικά η εντροπία Fuzzy χαρακτηρίζεται από την αβεβαιότητα, σε αντίθεση με την παραπάνω άποψη είναι αποτέλεσμα της ανεπάρκειας πληροφοριών γιατί δεν μπορούσαν να προβλεφθούν ακριβώς οι τιμές. Η κατανομή πιθανότητας κάποιας Fuzzy μεταβλητής εισάγεται με τη βοήθεια μίας συνάρτησης συμμετοχής, μ(x) ενός κανονικού συνόλου Fuzzy. Η συνάρτηση συμμετοχής παριστάνει αβεβαιότητα στο παρόν για ένα γεγονός που περιγράφεται ως Fuzzy σύνολο, ενώ η κατανομή πιθανότητας δείχνει την πιθανότητα που υπάρχει να συμβεί ένα γεγονός που

[31] Kosko (1986)
[32] Pal and Bezdek (1995)
[33] Li και Liu (2006) [2.39]

εκφράζεται με τιμές μίας Fuzzy μεταβλητής. Τελικά, από τους Li και Liu ορίστηκε η συνάρτηση:

$$S(t) = -t \ln t - (1-t)\ln(1-t)$$

με την υπόθεση ότι $0 \ln 0 = 0$. Η S(t) είναι κοίλη στο [0 ,1] και επίσης είναι συμμετρική στην τιμή t = 0,5.

Εάν ξ μία διακριτή Fuzzy μεταβλητή, που παίρνει τιμές από το διάστημα $\{x_1, x_2,..\}$,η εντροπία της μας δίνεται από τον τύπο:

$$\mathrm{H}(\xi) = \sum_{\iota=1}^{\infty} \mathrm{S}(\mathrm{Cr}\{\xi = \mathrm{x_i}\}) \qquad [2.40]$$

ενώ όταν η ξ είναι μία συνεχής Fuzzy μεταβλητή, τότε έχουμε:

$$\mathrm{H}(\xi) = \int_{-\infty}^{\infty} \mathrm{S}(\mathrm{Cr}\{\xi = \mathrm{r}\})\mathrm{dr} \qquad [2.41]$$

όπου:

$$\mathrm{S(t)} = -\mathrm{t}\ln\mathrm{t} - (1-\mathrm{t})\ln(1-\mathrm{t}) \qquad [2.42]$$

2.3.6 Η Generalized Εντροπία

Στην επιστήμη των μαθηματικών η generalized εντροπία ονομάζεται και f – divergence. Ένα είδος f–divergence είναι και η Kullback – Cross entropy divergence. Οι Friedrich Liese και Igor Vajda[34] γενίκευσαν την ποικίλη ολική απόσταση, την Hellinger divergence, την Kullback – Cross εντροπία και την Pearson divergence. Κατά την έρευνά τους οι βασικές ιδιότητες των f – divergences σχετικά με τα λάθη αποφάσεως αποδεικνύονται με άλλο τρόπο, δηλαδή με την αντικατάσταση της κλασικής ανισότητας του Jensen με μία επέκταση της μεθόδου Taylor των κυρτών συναρτήσεων. Η μέθοδος αυτή δείχνει εύκολα ότι οι f – divergences είναι οι μέσοι όροι των στατιστικών πληροφοριών οι οποίοι διαφέρουν στα σταθμά και μόνο, από τις προηγούμενες κατανομές.

Ο Rényi εισήγαγε μια σειρά μέτρων divergence κατανομών P, Q που είχαν τις ίδιες ιδιότητες με την Kullback-Cross entropy, την οποία περιλάμβαναν σαν μία ειδική περίπτωση.

[34] Liese και Vajda (2006)

Οι Ali και Silvey[35] ήταν οι ερευνητές οι οποίοι εισήγαγαν την f – divergence. Ο τύπος της είναι ο εξής:

$$D_f(P,Q) = \int \frac{dQ}{d\mu} f\left(\frac{\frac{dP}{d\mu}}{\frac{dQ}{d\mu}}\right) d\mu$$

Και αφορούσε τις κυρτές συναρτήσεις $f : (0,\infty) \to \mathscr{R}$, με μ να είναι ένα σ – διακριτό μέτρο το οποίο κυριαρχεί στα P, Q .

Όταν έχουμε $f(t) = t \ln t$ η f – divergence είναι ουσιαστικά η Kullback-Cross entropy divergence.

Σήμερα, λόγω της τεράστιας σημασίας των divergences στη θεωρία πληροφοριών, στη θεωρία πιθανοτήτων και στη στατιστική, μεγάλο μέρος της επιστημονικής κοινότητας έχει ενδιαφερθεί για την επέκταση και απλοποίηση της γενικής θεωρίας των f – divergences.

2.3.7 Η Εντροπία των Havrda – Charvát

Οι Havrda και Charvat[36] θεωρώντας ένα σύνολο B μη κενό και μία διαμέρηση $R(B)=\{M_{i....}M_N, \mu_{i.....}\mu_N\}$, στην οποία κάθε στοιχείο $M_i \in R(B)$ έχει μέτρο $\mu(M_i)$ όπου i=1,2,..,N. Ονόμασαν τη σχέση $S(\mu_1, \dots, \mu_v; a)$ structural α-entrop , εάν και μόνο εάν ισχύουν τα παρακάτω:

1) $S(\mu_1, \dots, \mu_N; a)$ είναι συνεχής στην περιοχή $\mu_i \geq 0$ και $\sum_{i=1}^{N} \mu_i = 1$ με $a > 0$
2) $S(1; a) = 0$ και $S\left(\frac{1}{2}, \frac{1}{2}; a\right) = 1$
3) $S(\mu_1, \dots, \mu_{i-1}, 0, \mu_{i+1}, \dots, \mu_N; a) = S(\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_N; a)$ για κάθε i=1,2,…,N
4) $S(\mu_1, \dots, \mu_{i-1}, \nu_{i1}, \nu_{i2}, \mu_{i+1}, \dots, \mu_N; a) = S(\mu_1, \dots, \mu_{i-1}, \mu_i, \mu_{i+1}, \dots, \mu_N; a) + a\mu_i^a S(\frac{\nu_{i1}}{\mu_i}, \frac{\nu_{i2}}{\mu_i}; a)$ για κάθε $\nu_{i1} + \nu_{i2} = \mu_i > 0$ με $i = 1,2, \dots, N$ και $a > 0$

Η παράμετρος α ονομάζεται χαρακτηριστική παράμετρος. Οι Havrda και Charvat χρησιμοποίησαν τις παραπάνω ισότητες για να αποδείξουν τα εξής:

1) $a = 1$

[35] Ali και Silvey (1966)
[36] Havrda και Charvat (1967)

2) Αν $\nu_k \geq 0, k = 1, \dots, m$ και $\sum_{k=1}^{m} \nu_k = \mu_i > 0,$ τότε $S(\mu_1, \dots, \mu_{i-1}, \nu_1, \dots, \nu_m, \mu_{i+1}, \dots, \mu_N; a) = S(\mu_1, \dots, \mu_N; a) + \mu_i^a S(\frac{\nu_1}{\mu_i}, \dots, \frac{\nu_m}{\mu_i}; a)$

3) Αν $\nu_{if} \geq 0, j = 1,2, \dots, m_i, \sum_{j=1}^{m_i} \nu_{ij} = \mu_i > 0\,, i = 1,2, \dots, n$ και $\sum_{i=1}^{i} \mu_i = 1$ τότε $S(\nu_{11}, \dots, \nu_{1m_1}, \dots, \nu_{n_1}, \dots, \nu_{nm_n}; a) = S(\mu_1, \dots, \mu_N; a) + \sum_{i=1}^{n} \mu_i^a\, S(\frac{\nu_{i1}}{\mu_i}, \dots, \frac{\nu_{im_i}}{\mu_i}; a)$

4) Αν $F(n, a) = S\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}; a\right)$ τότε $F(mn, a) = F(m, a) + \frac{1}{m^{a-1}} F(n, a) = F(n, a) + \frac{1}{n^{a-1}} F(m, a)$ για κάθε θετική σταθερά m,n

5) Αν $a \neq 1$ τότε $F(n, a) = c(a)(1 - \frac{1}{n^{a-1}})$, όπου $c(a)$ είναι συνάρτηση της χαρακτηριστικής παραμέτρου a κατέληξαν στα εξής:

$$S(\mu_1, \dots, \mu_N; a) = \frac{2^{a-1}}{2^{a-1} - 1}(1 - \sum_{i-1}^{N} \mu_i^a \;\; \text{για } a > 0 \; \textit{και} \; a \neq 1 \qquad [2.44]$$

και

$$S(\mu_1, \dots, \mu_N; 1) = -\sum_{i=1}^{N} \mu_i \log \mu_i \qquad [2.44]$$

2.3.8 Εντροπία Kapur[37]

Ο Kapur όρισε το μέτρο της εντροπίας τάξης α και μέτρου β ως:

$$H_{\alpha,\beta}(X) = \frac{1}{\beta - \alpha} \ln\left(\frac{\int (f_X(x))^a dx}{\int (f_X(x))^\beta dx}\right), \qquad \alpha \neq \beta, \qquad \alpha > 0, \qquad \beta > 0 \qquad [2.45]$$

2.3.9 Εντροπία Varma[38]

Ο Varma επίσης πρότεινε την γενικευμένη εντροπία τάξης α και τύπου β ως:

$$H_\alpha^\beta(X) = \frac{1}{\beta - \alpha} \ln\left(\int (f_X(x))^{\alpha+\beta-1} dx\right), \qquad \beta - 1 < \alpha < \beta, \qquad \beta \geq 1 \qquad [2.46]$$

[37] Kapur (1972)
[38] Varma (1966)

2.3.10 Δεσμευμένη Εντροπία

Δεσμευμένη είναι η εντροπία μίας τυχαίας μεταβλητής, εάν γνωρίζουμε μία άλλη τυχαία μεταβλητή. Σε αυτό το είδος εντροπίας ισχύουν τα παρακάτω:

$$= -\sum_{j=1}^{m} p\{y_i\} \sum_{k=1}^{n} p\{x_k|y_i\} \log p\{x_k|y_i\} = -\sum_{j=1}^{m}\sum_{k=1}^{n} p\{y_i\} p\{x_k|y_i\} \log p\{x_k|y_i\}$$

$$= -\sum_{j=1}^{m}\sum_{k=1}^{n} p\{y_i\} \frac{p\{x_k, y_i}{p\{y_i\}} \log p\{x_k|y_i\} = -\sum_{j=1}^{m}\sum_{k=1}^{n} p\{x_k, y_i\} \log p\{x_k|y_i\} \quad [2.48]$$

Κανόνας αλυσίδας:

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2/X_1) + \cdots + H(X_n/X_1X_2..X_{n-1}) \quad [2.49]$$

2.3.11 Διαφορική εντροπία

Η διαφορική εντροπία είναι μία ποσότητα που μοιάζει αλλά δεν είναι εντροπία. Έχει εφαρμογή σε πηγές διακριτού χρόνου με συνεχές αλφάβητο, στις οποίες οι έξοδοι είναι πραγματικοί αριθμοί[39].

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \log[f_X(x)] dx \quad [2.50]$$

Η διαφορική εντροπία δε χαρακτηρίζεται από το διαισθητικό νόημα της εντροπίας. Για να έχουμε αξιόπιστα την έξοδο μιας συνεχούς πηγής, είναι απαραίτητος ένας άπειρος αριθμός bits για κάθε έξοδο. Έχουμε δύο τύπους διαφορικής εντροπίας:

α) την από κοινού διαφορική εντροπία για δύο τυχαίες μεταβλητές:

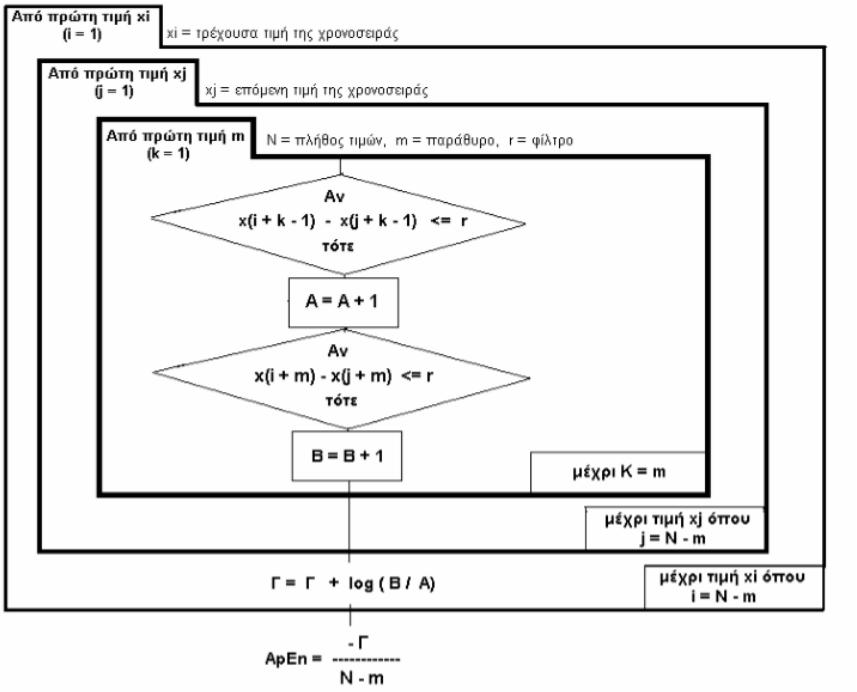
$$h(X, Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y) \log\left[f_{X,Y}(x, y)\right] dxdy \quad [2.51]$$

β) την υπό συνθήκη διαφορική εντροπία:

$$h(Y|X) = h(X, Y) - h(Y) \quad [2.52]$$

[39] Michalowicz και λοιποί (2008)

2.3.12 Προσεγγιστική εντροπία (ApEn)



Εικόνα 2.4. Λογικό διάγραμμα του αλγόριθμου της ApEn

Ο Pincus δημοσίευσε για πρώτη φορά τον αλγόριθμο της προσεγγιστικής εντροπίας (ApEn) στο άρθρο του «*A regularity statistic for medical data analysis*» το 1991. Ο συγκεκριμένος αλγόριθμος απαιτεί πολύπλοκους υπολογισμούς και γι' αυτό τον λόγο η ApEn είναι συνυφασμένη με τη χρήση λογισμικού. Η ApEn χρησιμοποιείται στην πράξη για:

- Τη σύγκριση της κανονικότητας (πολυπλοκότητας ή τυχαιότητας) δύο ή περισσότερων χρονοσειρών.
- Τη μελέτη της μεταβολής της πολυπλοκότητας σε σχέση με το χρόνο μέσα στην ίδια χρονοσειρά.

Τα τελευταία χρόνια η προσεγγιστική εντροπία βρίσκει εφαρμογή σε πολλούς κλάδους των βιοϊατρικών επιστημών με διάφορες παραλλαγές του αρχικού αλγορίθμου. Μια από τις σημαντικότερες είναι η cross-ApEn, η οποία συγκρίνει τη πολυπλοκότητα δύο χρονοσειρών προς την πολυπλοκότητα δύο άλλων.

Η ApEn είναι θετικός αριθμός, της οποίας ο πλήρης συμβολισμός είναι ApEn (m,r,N). Ο συμβολισμός αυτός προέρχεται από τις παραμέτρους N,m και r(1,2,3) από τις οποίες εξαρτάται άμεσα η τιμή ApEn. Ο ρόλος των συγκεκριμένων παραμέτρων είναι ο ακόλουθος:

- N: «Μήκος». Είναι το πλήθος (ακέραιος αριθμός) των τιμών που απαρτίζουν τη χρονοσειρά. Για αξιόπιστη τιμή ApEn το μέγεθος των χρονοσειρών (N) πρέπει να είναι μεγαλύτερο από 60 τιμές.
- m: «Παράθυρο». Είναι το πλήθος (ακέραιος αριθμός) των γειτονικών τιμών με τις οποίες συγκρίνεται κάθε τιμή της χρονοσειράς και λαμβάνει την τιμή 1 για σειρές 60-500 τιμές. Π.χ. αν m=2 κάθε τιμή συγκρίνεται με τις δυο επόμενες της, υπάρχουν δηλαδή δυο ζευγάρια σύγκρισης. Αν m=3 κάθε τιμή θα συγκριθεί με τρεις επόμενες κ.α.
- r: «Φίλτρο». Είναι ένας θετικός πραγματικός αριθμός που μετριέται σε μονάδες τυπικής απόκλισης (SD). Η τιμή του r κυμαίνεται από 0,15 SD έως 0,25 SD. Συνηθέστερη τιμή είναι το r=0,2 SD. Η τιμή r καθορίζει το πώς η τιμή ApEn θα επηρεαστεί από τον «θόρυβο» της χρονοσειράς. Λέγοντας θόρυβο ονομάζουμε τις μικρές διακυμάνσεις των τιμών της χρονοσειράς που εξαρτώνται από τη φυσιολογία του μελετούμενου συστήματος. Μεγάλη τιμή του r δίνει τιμή ApEn ανεπηρέαστη από τον θόρυβο. Μερικοί συγγραφείς προτείνουν η τιμή του r να είναι τρεις φορές μεγαλύτερη από το μέσο μέγεθος του αναμενόμενου θορύβου.

Μια τιμή ApEn από μόνη της δεν έχει καμία πρακτική σημασία αφού εξαρτάται από τις τιμές m,r,N. Δεδομένου ότι ο υπολογισμός της ApEn χρησιμοποιείται για τη σύγκριση συστημάτων θα πρέπει στις συγκρίσεις οι τιμές m,r και N να είναι κοινές για όλες τις συγκρινόμενες χρονοσειρές. Επιπλέον τα αριθμητικά δεδομένα που χρησιμοποιούνται για τον υπολογισμό της ApEn πρέπει να λαμβάνονται σε ίσα χρονικά διαστήματα.

2.3.13 Διαφορική Εντροπία Γκαουσιανής Τυχαίας Μεταβλητής

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad [2.53]$$

[2.54]

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx - \int_{-\infty}^{\infty} f(x) \log \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

$$= -\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2} \log e \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sigma^2 \log e = \frac{1}{2} \log(2\pi e\sigma^2)$$

2.4 **Ιδιότητες Πληροφορίας-Εντροπίας**

Οι ιδιότητες οι οποίες χαρακτηρίζουν τη μέση πληροφορία, όπως έχουν τεθεί και προκύπτουν από τον ορισμό της, και κυρίως από την κατά Shannon θεωρία αλλά και από άλλους ερευνητές, είναι οι παρακάτω[40]:

- Η μέση πληροφορία είναι συνεχής στο p, και κυρτή, κάτι που μπορούμε να το παρατηρήσουμε και από τη γραφική της συνάρτηση.
- Η μέση πληροφορία H(X) είναι επίσης συμμετρική, δηλαδή δεν επηρεάζεται από τη διάταξη των πιθανοτήτων. Έτσι, τυχαίες διαφορετικές μεταβλητές που έχουν κατανομές πιθανοτήτων προερχόμενες από την ίδια κατανομή πιθανοτήτων παρατηρούμε πως έχουν ίση εντροπία. Μάλιστα, σε μερικές περιπτώσεις, είναι πιθανόν και διαφορετικές κατανομές πιθανοτήτων να έχουν την ίδια μέση ποσότητα πληροφορίας.
- Η εντροπία H(X) έχει τη μέγιστη τιμή της στην περίπτωση που όλα τα ενδεχόμενα της είναι ισοπίθανα.
- Η εντροπία, όταν δύο τυχαίες ανεξάρτητες μεταβλητές X και Y συνδυάζονται, είναι προσθετική (additive).

2.5 **Αρχές Εντροπίας**

2.5.1 **Η Αρχή της Μέγιστης Εντροπίας του Jayne**

Τις περασμένες δεκαετίες, χρησιμοποιήθηκε η αρχή της μέγιστης εντροπίας του Jayne (Jayne's Maximum Entropy Principle ή MEP) για τη λύση μίας μεγάλης κατηγορίας πιθανολογικών συστημάτων. Η αρχή αυτή απαιτεί συγκέντρωση εννοιών της θεωρίας

[40] Ζορκάδης (2002)

πληροφοριών, βελτιστοποίηση και ακριβή γνώση της μερικής πληροφόρησης που μπορεί να κατέχει ένα άτομο αναφορικά με ένα πιθανολογικό σύστημα υπό το πρίσμα ενός συνόλου στατιστικών στιγμών. Το βασικό στοιχείο της αρχής της Μέγιστης Εντροπίας του Jayne, είναι ότι η πιο αμερόληπτη κατανομή πιθανοτήτων η οποία ικανοποιεί και όλους τους περιορισμούς είναι η maximum entropy κατανομή[41]. Στη μέγιστη εντροπία μπορούμε καταλήγουμε αν μεγιστοποιήσουμε το μέτρο της εντροπίας (συνήθως της εντροπίας του Shannon) κάτω από κάποιους περιορισμούς, με τη βοήθεια των πολλαπλασιαστών Lagrange. Αυτή η αρχή είναι γενικά αποδεκτή από όλη την επιστημονική κοινότητα τόσο για τα φιλοσοφικά της θεμέλια όσο και για τη μεγάλή επιτυχία της σε πρακτικές εφαρμογές. Λαμβάνοντας υπόψη το μέτρο της εντροπίας, την κατανομή πιθανοτήτων και το σύνολο των περιορισμών, η Jayne's Maximum Entropy Principle ή MEP παρέχει μία μεθοδολογία με στόχο τον προσδιορισμό της πιο αμερόληπτης κατανομής πιθανότητας.[42] Κατά την αρχή του Laplace, η πιο αμερόληπτη κατανομή, στην περίπτωση που κάποιος δεν είναι κάτοχος κάποιας προηγούμενης γνώσης, σχετικά με ένα πιθανολογικό συμβάν, είναι η ομοιόμορφη κατανομή. Συνεπώς, η MEP αποσκοπεί στον καθορισμό της πιο αμερόληπτης κατανομής πιθανότητας εάν τα δεδομένα είναι η συνάρτηση της εντροπίας του Shannon σε συνδυασμό με κάποιους απλούς γραμμικούς περιορισμούς. Η γραφική παράσταση αυτής της συνάρτησης είναι μία κοίλη συνάρτηση και άρα έχει ολικό μέγιστο. Η μεγιστοποίηση αυτή οδηγεί σε ένα σύνολο πιθανοτήτων, που δεν εμφανίζονται αρνητικές τιμές.

2.5.2 Η Αρχή της Ελάχιστης Cross-Εντροπίας του Kullback

Οι Kullback και Leibler[43] ήταν αυτοί που εισήγαγαν το μέτρο της εντροπίας απόστασης. Είναι μία μαθηματική μέθοδος που βασίζεται:

i. στο Kullback-Leibler(K-L) μέτρο, το οποίο είναι ένα μέτρο της απόστασης μεταξύ 2 πιθανοτικών κατανομών και εκφράζεται σαν το άθροισμα 2 λογαριθμικών όρων, με έναν από αυτούς να είναι οπωσδήποτε η μετρική Shannon και
ii. στην ελαχιστοποίηση του Kullback-Leibler (K-L) μέτρου, κάτω από συγκεκριμένους γραμμικούς περιορισμούς .

[41] Kesavan (2009)
[42] Favretti (2018)
[43] Kullback και Leibler (1951)

Η αρχή ελάχιστης εντροπίας ισοδυναμεί με την αρχή μέγιστης εντροπίας, όταν συντρέχει η μία κατανομή να είναι η ομοιόμορφη.

2.6 **Κωδικοποίηση**

2.6.1 Κωδικοποίηση Εντροπίας

Γίνεται περιορισμός των ακολουθιών χαρακτήρων που επαναλαμβάνονται. Είναι μία απλή τεχνική συμπίεσης που στηρίζεται στον εντοπισμό διαδοχικών επαναλήψεων του ίδιου χαρακτήρα και εν συνεχεία στην αντικατάστασή τους με τον αριθμό του πλήθους της εμφάνισής τους.

2.6.2 Στατιστική Κωδικοποίηση-Αντικατάσταση Προτύπων

Αναζητείται και στη συνέχεια μετριέται το πλήθος των ακολουθιών (χρωμάτων κλπ.). Ακολούθως, κάθε ακολουθία αντικαθίσταται από ένα μοναδικό κωδικό αριθμό (ένα ζευγάρι ειδικών χαρακτήρων= ένα byte)

2.6.3 Κωδικοποίηση Πηγής[44]

Σύμφωνα με τα όσα ορίζονται από το επικοινωνιακό μοντέλο, μετά από τη δημιουργία μηνυμάτων που πραγματοποιούνται από την πηγή, πραγματοποιείται η απαλοιφή δεδομένων για την όσο το δυνατόν απομάκρυνση κάθε πληροφορίας που δεν είναι σχετική με τον προορισμό. Όλα τα παραπάνω ισχύουν με τη υπόθεση ότι η πηγή παράγει μηνύματα τα οποία ενδιαφέρουν τον προορισμό στο σύνολό τους.

Για την καλύτερη απόδοση των επικοινωνιακών συστημάτων, απαιτείται η αφαίρεση οποιουδήποτε στοιχείου που θεωρείται πλεονασμός και εμπεριέχεται σε αυτά, με σκοπό την όσο το δυνατόν πιο συμπυκνωμένη αναπαράσταση των μηνυμάτων. Η παραπάνω διαδικασία καλείται κωδικοποίηση πηγής. Για το γενικό και το λεπτομερές επικοινωνιακό μοντέλο (Κεφάλαιο 1) σε πολλά σημεία του οποίου αντί του όρου κωδικοποίηση πηγής, χρησιμοποιήθηκε ο όρος συμπίεση.

Δίνοντας ένα πιο αναλυτικό ορισμό για την κωδικοποίηση πηγής, θα λέγαμε πως είναι η διαδικασία κατά την οποία οι ακολουθίες συμβόλων που παράγονται από την πηγή μετατρέπονται σε ακολουθίες συμβόλων κάποιου κώδικα (συνηθέστατα δυαδικές ακολουθίες), προκειμένου να αφαιρεθεί ο πλεονασμός και να προκύψει μία πιο συμπιεσμένη

[44] Ζορκάδης (2002)

απεικόνιση των μηνυμάτων. Δεδομένου ότι εξετάζουμε πηγές χωρίς μνήμη, (ανεξάρτητες ακολουθίες συμβόλων), αυτό που μας ενδιαφέρει κυρίως στη διαδικασία της κωδικοποίησης είναι πρωτίστως τα σύμβολα και όχι τόσο αυτά κάθε αυτά τα μηνύματα. Με τον τρόπο με τον οποίο θα πετύχουμε τη μετατροπή των συμβόλων πηγής σε ακολουθίες κωδικών συμβόλων, θα πετυχαίναμε και τη μετατροπή των μηνυμάτων της πηγής αντίστοιχα σε ακολουθίες κωδικών συμβόλων.

Για να μετατρέψουμε τα σύμβολα ή τις ακολουθίες συμβόλων της πηγής σε ακολουθίες κωδικών συμβόλων χρησιμοποιούμε διαφορετικά κωδικά σύμβολα που ονομάζονται ωδικό αλφάβητο. Επειδή όπως γνωρίζουμε, το δυαδικό αλφάβητο έχει, τα σύμβολα 0 και 1, η κωδικοποίηση, στην περίπτωση αυτή οδηγεί στην μετατροπή των συμβόλων ή των ακολουθιών συμβόλων σε δυαδικές ακολουθίες. Σε αυτή την κωδικοποίηση, το κάθε παραγόμενο από την πηγή σύμβολο, παριστάνεται με μια κωδική λέξη η οποία αποτελείται από ακολουθία κωδικών συμβόλων. Κώδικας, λοιπόν, ονομάζεται ένα σύνολο κωδικών λέξεων και η αντιστοίχιση τους με σύμβολα της πηγής ή ακολουθίες συμβόλων στην περίπτωση που πρόκειται για κωδικοποίηση μηνυμάτων.

2.6.4 Κωδικοποίηση Huffman

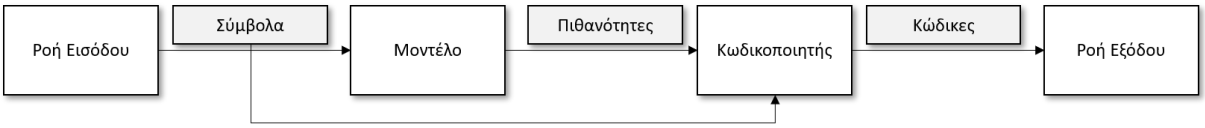
Η Κωδικοποίηση Huffman είναι βασισμένη στη στατιστική ανάλυση των δεδομένων. Με τη βοήθεια του αλγόριθμου κωδικοποίησης Huffman, ο οποίος είναι πολύ απλός προκύπτει ένας καθόλα άρτιος κώδικας για δεδομένες πιθανότητες εμφάνισης των συμβόλων της πηγής.[45] Εύκολα αποδεικνύεται ότι δεν υπάρχει άλλος που να μπορεί να οδηγήσει στη δημιουργία κώδικα που να έχει μικρότερο μέσο μήκος κωδικών λέξεων για ένα δεδομένο αλφάβητο πηγής.

Η δυαδική κωδικοποίηση των συμβόλων της πηγή, σύμφωνα με τον αλγόριθμο του Huffman, ακολουθεί τα εξής βήματα:

1. Αρχικά, διατάσσονται κατά φθίνουσα πιθανότητα εκπομπής, τα σύμβολα της πηγής.
2. Ενώνονται σε ένα, τα δύο τελευταία σύμβολα που έχουν τη μικρότερη πιθανότητα παραγωγής, με πιθανότητα η οποία είναι ίση με το άθροισα των πιθανοτήτων των δύο τελευταίων αυτών συμβόλων. Σαν αποτέλεσμα έχουμε τη μείωση του πλήθους των συμβόλων που αποτελούν το αλφάβητο της πηγής, κατά ένα.

[45] Thomas και Cover (1991)

3. Τα παραπάνω δύο βήματα, επαναλαμβάνονται μέχρι που να φτάσουμε το αλφάβητο της πηγής να αποτελείται από δύο μόνο σύμβολα. Αυτά τα δύο σύμβολα απεικονίζονται με τα 0 και 1 του δυαδικού κώδικα.
4. Ένα «0» και ένα «1» παριστάνουν , αντίστοιχα τα δύο σύμβολα τα οποία στο βήμα 2 είχαν συγχωνευτεί σε ένα. Το βήμα αυτό σχετίζεται με όλες τις συγχωνεύσεις.
5. Οι κωδικές λέξεις των συμβόλων σχηματίζονται από όλα τα ψηφία «0» και «1» που σχετίζονται μ' αυτά τα σύμβολα, δηλαδή απ' όλα τα ψηφία που αποδόθηκαν απευθείας σ' αυτά ή στα συγχωνευμένα σύμβολα στα οποία συμμετέχουν.



Εικόνα 2.5. Κωδικοποίηση Huffma

3 Εντροπία & Εφαρμογές

3.1 Η έννοια του Προβλήματος

Πρόβλημα ονομάζεται μια κατάσταση η οποία χρήζει αντιμετώπισης και απαιτεί μία μη προφανή και γνωστή λύση. Επίλυση, λύση ή αντιμετώπιση του προβλήματος είναι ο τρόπος με τον οποίο βρίσκει κάποιος το ζητούμενο, δηλαδή πετυχαίνει τον επιθυμητό στόχο. Η επίλυση ενός προβλήματος περιλαμβάνει ορισμένα στάδια, τα οποία αλληλεπιδρούν μεταξύ τους[46]:

1. Κατανόηση ενός προβλήματος, η οποία κατανόηση απαιτεί τη σωστή και πλήρη αποσαφήνιση τόσο των δεδομένων όσο και των ζητουμένων του προβλήματος.
2. Ανάλυση, δηλαδή τμηματοποίηση του προβλήματος σε απλούστερα υπό-προβλήματα, τα οποία να επιλύνονται πιο εύκολα.
3. Επίλυση μέσω της λύσης των επιμέρους προβλημάτων στα οποία έχει διαχωριστεί κατά την ανάλυση.

Υπολογιστικό πρόβλημα είναι ένα πρόβλημα το οποίο ένας υπολογιστής είναι ικανός να λύσει.

3.1.1 Προτυποποίηση Προβλημάτων

Ένα πρότυπο επιχειρησιακής έρευνας είναι αυτό που καθορίζει μια εξιδανικευμένη αναπαράσταση μέσα σε ένα πραγματικό σύστημα. Το σύστημα αυτό μπορεί να υπάρχει αλλά μπορεί και να είναι απλά μια ιδέα που πρέπει να μορφοποιηθεί για να γίνει πραγματικότητα. Στην πρώτη περίπτωση ο κύριος στόχος είναι η ανάλυση της συμπεριφοράς ενός συστήματος, με σκοπό τη βελτίωση της απόδοσής του, ενώ στη δεύτερη περίπτωση ο στόχος είναι πρωτίστως η βέλτιστη δομή του συστήματος που τελεί υπό ανάπτυξη. Ένα πραγματικό σύστημα είναι πολύπλοκο λόγω του μεγάλου αριθμού των μεταβλητών που επηρεάζουν τη συμπεριφορά του συστήματος. Το πραγματικό σύστημα απλοποιείται με τη χρήση ενός προτύπου αφού αρχικά καθοριστούν οι κυρίαρχες μεταβλητές και οι σχέσεις που τις διέπουν.

Η αναγνώριση προτύπων στην πληροφορική δεν είναι κάτι εύκολο, όπως είναι για τους ζώντες οργανισμούς και βασικά για τον άνθρωπο. Ένας ηλεκτρονικός υπολογιστής, πρέπει να «εκπαιδευθεί» κατάλληλα προκειμένου να αναγνωρίζει και να κατηγοριοποιεί τα πρότυπα

[46] Barnes και λοιποί (1997)

αυτόματα σε κατηγορίες. Ανάλογα με την εφαρμογή γίνεται κατάταξη των αντικειμένων σε κλάσεις με τη βοήθεια αλγορίθμων ταξινόμησης.

Γι' αυτό το λόγο, η αναγνώριση προτύπων σ' αυτόν τον τομέα αποτελεί ιδιαίτερο επιστημονικό πεδίο που έχει σαν στόχο την ανάπτυξη αλγορίθμων με τη βοήθεια των οποίων θα πραγματοποιείται μία αυτοματοποιημένη απόδοση κάποιου διακριτικού στοιχείου ή κάποιας τιμής σε εισαγόμενα δεδομένα, κατά κύριο λόγο κωδικοποιημένα σε αλληλουχίες αριθμών. Η μορφή των παραπάνω αλγόριθμων για την αναγνώριση προτύπων, εξαρτάται ιδιαίτερα από τον τύπο της εξόδου, αλλά και από το εάν ο αλγόριθμος είναι στατιστικά στη φύση ή όχι. Οι στατιστικοί αλγόριθμοι μπορούν να κατηγοριοποιηθούν περαιτέρω ως παραγωγικοί ή διακριτικοί. Με αυτόν τον τρόπο, τα δεδομένα ταξινομούνται αυτόματα σε ομάδες ή κατηγορίες με βάση ορισμένα κριτήρια, ακόμα και εάν η παρουσία κάποιου θορύβου δυσκολεύει την αναγνώριση, και ωθεί τα δεδομένα να μοιάζουν τυχαία σε μεγαλύτερο βαθμό από αυτό που είναι στην πραγματικότητα. Το ενδιαφέρον των ερευνητών για την αναγνώριση προτύπων το βρίσκουμε κάπου στη δεκαετία του 1960, με την ανάπτυξη της πληροφορικής και ειδικότερα, της τεχνητής νοημοσύνης.

Σε όλη τη διαδικασία αναγνώρισης προτύπων, το πρόβλημα που παρατηρείται είναι η κατάταξη νέων παρατηρήσεων ή αντικειμένων, σε υποομάδες που ήδη προϋπάρχουν ή σε ένα δεδομένο σύνολο αντικειμένων.

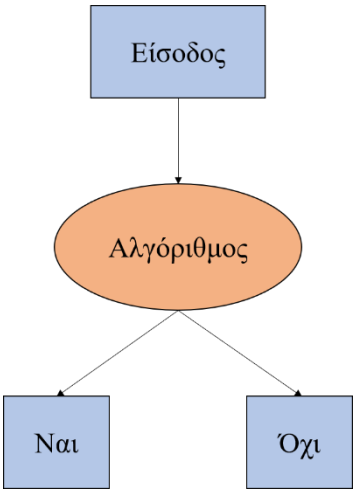
Ο πιο σημαντικός τύπος προτύπου προβλήματος είναι ο συμβολικός ή αλλιώς μαθηματικός τύπος προτύπου. Σε αυτά τα πρότυπα προβλήματα, υποθέτουμε ότι όλες οι συσχετιζόμενες μεταβλητές είναι ποσοτικές. Η βέλτιστη (optimal) λύση ενός πρότυπου προβλήματος, επιτυγχάνεται όταν οι τιμές των μεταβλητών απόφασης οδηγούν στη βέλτιστη τιμή της συνάρτησης, ικανοποιώντας ταυτόχρονα όλους τους περιορισμούς που έχουν τεθεί.

3.1.2 Προβλήματα Απόφασης

Σύμφωνα με τη θεωρία της υπολογισιμότητας και της πολυπλοκότητας, πρόβλημα απόφασης είναι ένα πρόβλημα σε κάποιο σύστημα, που επιδέχεται μια ναι ή όχι απάντηση, ανάλογα πάντα με τις τιμές κάποιων παραμέτρων εισόδου. Είναι με άλλα λόγια μία ερώτηση σε κάποιο άπειρο σύνολο εισόδων, η οποία απαντάται με «ναι» ή «όχι».

Οι είσοδοι μπορεί να είναι φυσικοί αριθμοί, αλλά μπορεί να είναι και τιμές άλλου είδους, π.χ. συμβολοσειρές στο δυαδικό αλφάβητο {0,1} ή άλλο σύνολο συμβόλων.

Είναι, γενικά, αποδεκτό ότι εάν χρησιμοποιήσουμε κάποια άλλη κωδικοποίηση όπως η Γκεντελοποίηση, τότε οποιαδήποτε συμβολοσειρά μπορεί να κωδικοποιηθεί και να αποδοθεί με κάποιον φυσικό αριθμό, κι έτσι στο τέλος το πρόβλημα απόφασης μπορεί να οριστεί και σαν υποσύνολο των φυσικών αριθμών.



Εικόνα 3.1. Πρόβλημα Απόφασης

Ένα πρόβλημα απόφασης που μπορεί να επιλυθεί με κάποιον αλγόριθμο, ονομάζεται αποκρίσιμο ή αποφασίσιμο.

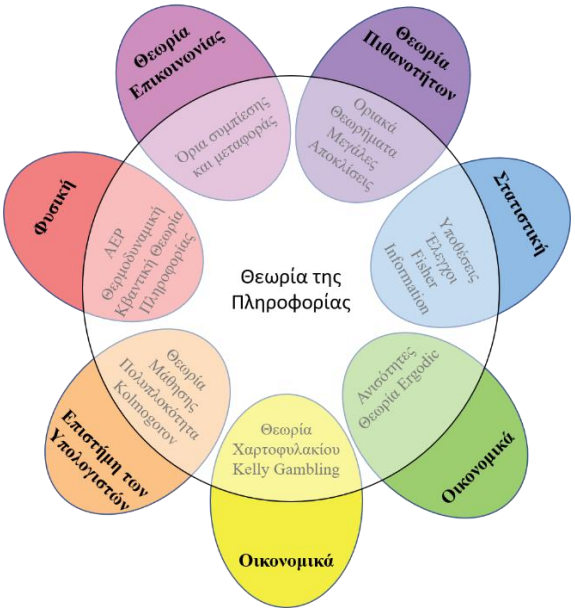
Μια μέθοδος επίλυσης ενός προβλήματος απόφασης, που δίνεται με τη μορφή αλγορίθμου, ονομάζεται διαδικασία απόφασης.

Τα προβλήματα απόφασης περιέχουν σημαντικούς δυσδιάκριτους παράγοντες οι οποίοι δεν μπορούν να εκφραστούν τέλεια σε όρους ενός μαθηματικού προτύπου. Ο σημαντικότερος δε από αυτούς είναι το ανθρώπινο στοιχείο, που έχει σημαντικό ρόλο σε οποιοδήποτε πρόβλημα απόφασης.

Σε μια επιχείρηση προβλήματα απόφασης μπορεί να τίθενται με ερωτήματα όπως για παράδειγμα αν θα παράξει η επιχείρηση ένα νέο προϊόν ή αν θα προσλάβει νέο προσωπικό.

3.2 **Εφαρμογές Εντροπίας**

Οι τομείς και οι θεωρίες στους οποίους έχει εφαρμογή ή επηρεάζει η εντροπία της πληροφορίας όσον αφορά τα προβλήματα απόφασης είναι πολλοί. Μερικοί από αυτούς είναι οι εξής: Οικονομία, Μαθηματικά & Στατιστική, Αρχιτεκτονική υπολογιστών, Εφαρμοσμένη μηχανική, Φυσική, Ιατρική, Νευρολογία, Θεωρία Παιγνίων, Θεωρία Μαύρης Τρύπας, Αλγοριθμική Θεωρία Πολυπλοκότητας, Πηγαίος Κώδικας και Χωρητικότητα καναλιού. Παρακάτω, αναλύονται οι σπουδαιότεροι από αυτούς τους τομείς.



Εικόνα 3.2. Σχέση της θεωρίας πληροφοριών με άλλους τομείς
Πηγή: Wiley (2006)-Elements of information theory[2]

3.2.1 Ασφάλεια Δικτύων Υπολογιστών

Το κυριότερο πρόβλημα του πεδίου ασφάλειας στα δίκτυα υπολογιστών είναι η ανίχνευση των επιθέσεων και η αντιμετώπισή τους. Οι κυβερνοεπιθέσεις, που είναι συχνό φαινόμενο της εποχής μας, θέτουν σε κίνδυνο την ασφάλεια των δικτύων. Η παραμετροποίηση της, δηλαδή τα χαρακτηριστικά (features) πακέτων, που συλλέγονται με την παρακολούθηση της κίνησης μέσα σ' ένα δίκτυο, βοηθά στην πιο αποδοτική ανίχνευση των επιθέσεων DDOS. Αποδοτική ανίχνευση είναι η ανίχνευση που αρχικά εντοπίζει με επιτυχία μία επίθεση και στη συνεχεία εξάγει τις όσο το δυνατόν λιγότερο λανθασμένες ενδείξεις (falsepositives, falsenegatives). Η μέθοδος της εντροπίας εφαρμόζεται σε δείγματα δεδομένων στα οποία υπάρχει «ομαλή» και «ανώμαλη» κίνηση υπηρεσίας και γίνεται προσπάθεια για τον προσδιορισμό ενός δείκτη που να διακρίνει τις ανωτέρω δύο κατηγορίες (classifier).

Για την ανίχνευση DDoS επιθέσεων εφαρμόζονται δύο κύριες προσεγγίσεις: α) η προσέγγιση Υπογραφής, γνωστή ως Signature Based Approach-SBA και β) η Προσέγγιση Ανωμαλιών, δηλαδή η Anomaly Based Approach-ABA. Τα χαρακτηριστικά των επιθέσεων της SBA τίθενται σε σύγκριση με μία βάση δεδομένων γνωστών κακόβουλων απειλών και υπογραφών. Αυτός ο τρόπος για την ανίχνευση κακόβουλου λογισμικού λειτουργεί περίπου όπως τα περισσότερα antivirus. Βασικό μειονέκτημα αυτής της μεθόδου είναι η καθυστέρηση που ανακύπτει όταν πρέπει να ενημερώσει τη αντίστοιχή βάση δεδομένων για την απειλή και την υπογραφή της. Σ' αυτό το διάστημα, λοιπόν, δε είναι δυνατή η ανίχνευση νέων απειλών καθώς επίσης είναι και αδύνατη η ανίχνευση επιθέσεων zero day - μηδενικής μέρας.

Τα παραπάνω προβλήματα επιχειρείται να αρθούν με την προσέγγιση ABA. Με τη μέθοδο αυτή γίνεται χρήση στατιστικών προσεγγίσεων και προσεγγίσεων ανάλυσης κατανομής και εξόρυξης δεδομένων.

Η ABA χρησιμοποιεί επιλεγμένους δείκτες με τη βοήθεια των οποίων θεωρεί και χαρακτηρίζει μία κίνηση ως «φυσιολογική» και τη χρησιμοποιεί σαν βάση σύγκρισης με την κίνηση του διαδικτύου. Τα κριτήρια που χρησιμοποιεί έχουν σχέση με το εύρος ζώνης, τα πρωτόκολλα, τις συσκευές που συνδέονται μεταξύ τους, τις θύρες κλπ. Σκοπός στην προσέγγιση αυτή είναι η ειδοποίηση του διαχειριστή του δικτύου κάθε φορά που εντοπίζεται μία «ανώμαλη» κυκλοφορία, δηλαδή μία κυκλοφορία που διαφέρει σημαντικά από την

κυκλοφορία πρότυπο, τη «φυσιολογική». Το μειονέκτημα σε αυτή τη μέθοδο, είναι ότι παρατηρούνται μεταξύ των άλλων και αρκετοί ψευδείς συναγερμοί.

Μία από τις κυριότερες μεθόδους ανίχνευσης επιθέσεων άρνησης υπηρεσιών είναι η στατιστικομαθηματική προσέγγιση της εντροπίας κατά Shannon. Η προσέγγιση ανίχνευσης επιθέσεων χρησιμοποιώντας την εντροπία βοηθάει σημαντικά στην ανίχνευση των επιθέσεων DDoS. Η εντροπία σ' αυτή την περίπτωση αποτελεί το μέτρο αταξίας του συστήματος που ερευνάται. Ο ορισμός μίας κίνησης ως «φυσιολογική» προϋποθέτει συνεχή παρακολούθηση της κίνησης του δικτύου. Η εφαρμογή της εντροπίας είναι αρκετά αποτελεσματική, απαιτεί όμως δίκτυο υψηλής ταχύτητας και μνήμη.

Η θεωρία της πληροφορίας σχετικά με την ανίχνευση ανωμαλιών αναλύεται από τους Lee και Xiang οι οποίοι εντρύφησαν στη χρήση μέτρων της θεωρίας πληροφορίας (εντροπία, υπό συνθήκη εντροπία, σχετική υπό συνθήκη εντροπία, κέρδος πληροφορίας, πληροφοριακό κόστος). Από τους ανωτέρω επιστήμονες καθορίστηκε ο τρόπος με τον οποίο αυτά τα μεγέθη λειτουργούν στην ανίχνευση ανωμαλιών και εν συνεχεία, πώς μπορούν να εφαρμοστούν στα διάφορα σύνολα δεδομένων (datasets).

Η εντροπία μπορεί να αποτελέσει για τα εξεταζόμενα δεδομένα μέτρο κανονικότητας που μας δείχνει την πιθανότητα να εμφανιστεί το ενδεχόμενο, που αντιστοιχεί σε κάποιο πεδίο της επικεφαλίδας, όπως διεύθυνση πηγής ή διεύθυνση προορισμού.

Με την υπό συνθήκη εντροπία περιγράφεται το ποσοστό αβεβαιότητας που μένει άν παρατηρηθεί ένα γεγονός ή ένα υποσύνολο γεγονότων που εξετάζονται . Κατά τους Lee και Xiang[47] η υπό συνθήκη εντροπία αποτελεί ως επί το πλείστων «μέτρο της κανονικότητας διαδοχικών εξαρτήσεων» και όσο μικρότερη είναι η τιμή της, τόσο το καλύτερο. Όμως, η μοντελοποίηση των συστημάτων με υψηλή υπό συνθήκη εντροπία είναι αρκετά δύσκολη.

Στις περιπτώσεις που η υπό συνθήκη εντροπία εφαρμόζεται σε σύνολο δεδομένων, με την υπό συνθήκη εντροπία μπορεί να υπολογιστεί η απόσταση μεταξύ δύο συνόλων δεδομένων. Συνεπώς, ανωμαλίες είναι εφικτό να ανιχνευθούν εάν η εντροπία του συστήματος συγκρίνεται ανά μικρά χρονικά διαστήματα όπως μέρα με τη μέρα (ή άλλη χρονική περίοδο) με βάση την εντροπία της προηγούμενης ημέρας.

[47] Lee και Xiang (2000)

Οι Nychis και λοιποί[48] αναλύοντας τις διαφορετικές τεχνικές με δείκτες της θεωρίας της πληροφορίας για ανίχνευση ανωμαλιών αναφορικά με την κίνηση δικτύου υπολογιστώ, δήλωσαν ότι: «Οι βασισμένες στην εντροπία προσεγγίσεις για την ανίχνευση ανωμαλιών παρουσιάζουν ιδιαίτερο ενδιαφέρον, δεδομένου ότι παρέχουν πιο λεπτομερή στοιχεία από την παραδοσιακή ανάλυση του όγκου της κίνησης» (Nychis και λοιποί 2008).

3.2.2 Η Προσεγγιστική Εντροπία στην Ιατρική

Ο Pincus[49] ξεκίνησε τις πρώτες εφαρμογές της προσεγγιστικής εντροπίας (ApEn), ακριβώς μετά το άρθρο του. Ο πρώτος κλάδος της ιατρικής όπου εφαρμόστηκε ο αλγόριθμος αυτός ήταν η καρδιολογία. Ειδικότερα, λόγω της εύκολης δημιουργίας μεγάλων χρονοσειρών από τις μετρήσεις καρδιογραφημάτων, από ιατρική και πρακτική πλευρά πολλές ομάδες ερευνητών άρχισαν να χρησιμοποιούν την ApEn μαζί με άλλα εργαλεία στατιστικής. Άλλοι κλάδοι της ιατρικής στους οποίους η ApEn βρίσκει ουσιαστική εφαρμογή είναι η ενδοκρινολογία όπως επίσης και η νευρολογία και συγκεκριμένα στην αξιολόγηση ηλεκτροεγκεφαλογραφημάτων.

3.2.3 Οικονομία και Μελέτη Χρηματιστηρίων

Μέσω του βασικότερου μέτρου της θεωρίας της πληροφορίας, της εντροπίας του Shannon, μπορεί να προβλεφθεί μία ενδεχόμενη κρίση. Πιο συγκεκριμένα, όταν οι τιμές της εντροπίας είναι υψηλές, δείχνουν αύξηση της αβεβαιότητας στο χώρο και εκπέμπεται σήμα κινδύνου κρίσης στο μέλλον. Όταν οι τιμές είναι πολύ χαμηλές, η οικονομική κρίση είναι προ των πυλών. Στην οικονομία η εντροπία της πληροφορίας αποτελεί μέτρο αποδόμησης και αποδιοργάνωσης του συστήματος ή όχι. Μία χαμηλή – ασύμμετρη πληροφόρηση, είναι δείκτης μεγάλης αβεβαιότητας και υψηλής δαπάνης της ενέργειας. Φανερώνει ένα σύστημα που δουλεύει περισσότερο από τις αντοχές του, με υψηλή φορολογία, δημοσιονομικό χρέος, υψηλά επιτόκια δανεισμού, μεγάλο ύψος ληξιπρόθεσμων οφειλών νοικοκυριών, χαμηλούς μισθούς, μεγάλη ανεργία, χαμηλές αποταμιεύσεις και χαμηλή κατανάλωση, πληθωρισμό τιμών, λίγες επενδύσεις κ.λπ. που ανεβάζουν σταδιακά την εντροπία της αγοράς. Όταν η εντροπία ανεβεί, όλα τα ενδεχόμενα είναι δυνατά (οικονομικές και κοινωνικές αναταραχές, σοκ–αντίδραση στο σοκ). Μέσω της εντροπίας της πληροφορίας οι κοινωνίες οι οποίες

[48] Nychis και λοιποί (2008)
[49] Pincus και λοιποί (1991)

ελέγχουν και διακινούν σωστά τη ροή πληροφορίας επιδιώκουν την ενδογενή ανάπτυξη και την επαναφορά στην συμμετρία και την ισορροπία.

3.2.4 Η εντροπία ως Μέτρο Κεφαλαιακής Αύξησης

Η incremental entropy δημιουργήθηκε από τον ερευνητή Ou[50] και μετράει το χρόνο που χρειάζεται για να διπλασιάσουμε το κεφάλαιό μας. Βάσει της incremental entropy μπορούμε να κάνουμε βελτιστοποίηση χαρτοφυλακίου. Ερευνητές έχουν βρει ότι η incremental entropy, ένα είδος γενικευμένης εντροπίας, θα μπορούσε να χρησιμοποιηθεί στη βελτιστοποίηση χαρτοφυλακίων. Η νέα αυτή θεωρία χαρτοφυλακίου έχει κάποιες από τις ιδιότητες της θεωρίας του Markowitz, αλλά τονίζει ότι η αυξητική ταχύτητα του κεφαλαίου είναι ένα πιο αντικειμενικό κριτήριο στην αξιολόγηση των χαρτοφυλακίων. Δεδομένων των προβλέψεων των αποδόσεων μπορεί να επιτευχθεί η βέλτιστη αναλογία της επένδυσης. Συνδυάζοντας, λοιπόν, τη νέα θεωρία χαρτοφυλακίου και τη γενική θεωρία πληροφοριών μπορούμε να προσεγγίσουμε ένα μέτρο το οποίο θα αντιπροσωπεύει αυτή την αυξητική ταχύτητα του κεφαλαίου, με την προϋπόθεση βέβαια ότι παρέχονται κάποιες πληροφορίες. Πιο γενικά, ο αριθμός των επενδύσεων είναι μεγαλύτερος από ένα και οι μελλοντικές τους αποδόσεις είναι αβέβαιες. Όταν δίνεται όμως η joint κατανομή πιθανότητας των μελλοντικών τους αποδόσεων, τότε πως μπορούμε να καθορίσουμε το βέλτιστο χαρτοφυλάκιο;

Ο Markowitz[51] επισημαίνει ότι ένα αποδοτικό χαρτοφυλάκιο είναι είτε ένα χαρτοφυλάκιο που δίνει τη μεγαλύτερη αναμενόμενη απόδοση για ένα δεδομένο επίπεδο κινδύνου ή ένα χαρτοφυλάκιο με το μικρότερο κίνδυνο για μία δεδομένη αναμενόμενη απόδοση. Όσο για τον τρόπο με τον οποίο μπορούμε να αξιολογήσουμε ένα χαρτοφυλάκιο δεδομένης μίας αναμενόμενης απόδοσης και μίας τυπικής απόκλισης δεν υπάρχει κάποιο αντικειμενικό κριτήριο. Προφανώς, λοιπόν, το αποδοτικό χαρτοφυλάκιο του Markowitz δεν είναι το χαρτοφυλάκιο που χρειαζόμαστε για την ταχύτερη αύξηση του κεφαλαίου ή για αποδοτικό χαρτοφυλάκιο, για αυτό και χρειαζόταν ένα νέο μαθηματικό μοντέλο.

Συγκρίνοντας την incremental entropy με τη θεωρία του Markowitz βλέπουμε ότι αυτή η νέα θεωρία στηρίζει τα συμπεράσματα του Markowitz ότι ο κίνδυνος επένδυσης μπορεί να μειωθεί από ένα αποδοτικό χαρτοφυλάκιο. Υπάρχουν όμως σημαντικές διαφορές:

[50] Ou (2005)
[51] Markowitz (1959)

- Η νέα θεωρία υιοθετεί την απόδοση του γεωμετρικού μέσου ως ένα αντικειμενικό κριτήριο για τη βελτιστοποίηση χαρτοφυλακίου και παρέχει μεθόδους για βελτιστοποίηση των επενδυτικών αναλογιών.
- Η νέα θεωρία ασχολείται με την επέκταση και την πιθανότητα ζημίας και κέρδους αντί της προσδοκίας και της τυπικής απόκλισης της απόδοσης για να περιγράψει την επενδυτική αξία.

3.2.5 Η θεωρία Παιγνίων

Στη Θεωρία Παιγνίων εξετάζονται προβλήματα απόφασης τα οποία έχουν να κάνουν με πολλούς συμμετέχοντες, όπου κάθε ένας συμμετέχων έχει τους δικούς του στόχους σχετικά με ένα κοινό σύστημα ή σχετικά με τον ανταγωνισμό που αναπτύσσεται για το μοίρασμα κοινών πόρων. Η Θεωρία Παιγνίων είναι δημιούργημα σεναρίων ανταγωνισμών και γι’ αυτό τα προβλήματα που εξετάζει είναι γνωστά σαν «παίγνια» και οι συμμετέχοντες σε αυτά είναι γνωστοί ως "παίκτες". Ο κάθε παίκτης έχει στη διάθεσή του ένα σύνολο «κινήσεων-στρατηγικών», τις οποίες δύναται να εκτελέσει. Κάθε κίνηση που κάνει προκαλεί και μια «απόδοση» για τον συγκεκριμένο παίκτη. Οι παίκτες επιλέγουν με τέτοιον τρόπο τις κινήσεις τους, ώστε να έχουν σαν αποτέλεσμα τη μεγιστοποίηση των κερδών τους, και αντιλαμβάνονται πολύ καλά ότι και όλοι οι υπόλοιποι παίκτες που συμμετέχουν στο παίγνιο επιδιώκουν το ίδιο με αυτόν, τη μεγιστοποίηση των αποδόσεών τους.

4 Εξόρυξη δεδομένων (Data Mining)

4.1 Το πρόβλημα της ταξινόμησης

Ο όρος Data Mining αναφέρεται στην εξαγωγή πολύτιμων γνώσεων από μεγάλους όγκους δεδομένων. Η εξόρυξη δεδομένων είναι η διαδικασία ανακάλυψης γνώσης από δεδομένα. Με τον τεράστιο όγκο δεδομένων που αποθηκεύονται, είναι πολύ σημαντικό να αναπτυχθούν εργαλεία ανάλυσης και λήψης αποφάσεων με σκοπό την εξαγωγή ενδιαφέρουσας γνώσης. Το καθήκον της ταξινόμησης είναι να ασχολείται με την πρόβλεψη της τιμής ενός πεδίου χρησιμοποιώντας τις τιμές ενός άλλου. Στη μηχανική μάθηση, η ταξινόμηση είναι ένα παράδειγμα εποπτευόμενης μάθησης. Ένα σύνολο δεδομένων με σωστά προσδιορισμένες παρατηρήσεις χρησιμοποιείται για την εκπαίδευση ενός αλγορίθμου. Η μη επιβλεπόμενη διαδικασία ονομάζεται ομαδοποίηση, στην οποία έχουμε ομαδοποίηση των δεδομένων σε κατηγορίες με βάση την ομοιότητα ή την απόσταση. Οι παρατηρήσεις αναλύονται σε ένα σύνολο μετρήσιμων ιδιοτήτων οι οποίες ονομάζονται χαρακτηριστικά. Ο αλγόριθμος που εφαρμόζει την ταξινόμηση ονομάζεται ταξινομητής.

Στη μηχανική μάθηση, οι παρατηρήσεις είναι γνωστές και ως *περιπτώσεις*, οι ερμηνευτικές μεταβλητές ως *χαρακτηριστικά* και οι κατηγορίες που είναι να προβλεφθούν ονομάζονται *κλάσεις*.

Η ταξινόμηση χρησιμοποιείται ως διαδικασία εξόρυξης δεδομένων καθώς και για πιο λεπτομερή στατιστική μοντελοποίηση και έχει πολλά πεδία εφαρμογών. Μερικά από αυτά είναι η οπτική αναγνώριση χαρακτηριστικών, η αναγνώριση ομιλίας, οι μηχανές αναζήτησης στο διαδίκτυο και η ιατρική διάγνωση.

Οι περισσότεροι υπάρχοντες αλγόριθμοι εξόρυξης (συμπεριλαμβανομένων αλγορίθμων ταξινόμησης, ομαδοποίησης, ανάλυσης συσχέτισης κ.λπ.) εργάζονται σε μεμονωμένους πίνακες. Για παράδειγμα, ένας τυπικός αλγόριθμος ταξινόμησης (π.χ. SVM) λειτουργεί σε έναν πίνακα που περιέχει πολλές πλειάδες, καθεμία από τις οποίες έχει μια ετικέτα κλάσης και μια τιμή σε κάθε χαρακτηριστικό του πίνακα. Τα τελευταία χρόνια υπάρχει αυξανόμενο ενδιαφέρον για πολυεθνική έρευνα και εφαρμογή ταξινόμησης, που υποδεικνύει την

δυσκολία αντιμετώπισης μεγάλων χώρων αναζήτησης σχέσεων, σύνθετες σχέσεις μεταξύ σχέσεων και ένας τρομακτικός αριθμός εμπλεκόμενων χαρακτηριστικών.[52]

4.1.1 Αλγόριθμοι ταξινόμησης

- K-Nearest Neighbors (KNN)[53]

Ο αλγόριθμος Κ πλησιέστερων γειτόνων (KNN) είναι ένας τύπος εποπτευόμενου αλγορίθμου ML που μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση. Στον κλάδο αυτό χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης.

Ακολουθούν κάποια σημαντικά στοιχεία για τον αλγόριθμο KNN

- Είναι ένας από τους απλούστερους αλγόριθμους μηχανικής μάθησης που βασίζονται στην τεχνική εποπτευόμενης μάθησης.
- Υποθέτει την ομοιότητα μεταξύ των δεδομένων και των διαθέσιμων περιπτώσεων και θέτει τη νέα περίπτωση στην κατηγορία που είναι πιο παρόμοια με τις διαθέσιμες κατηγορίες.
- Αποθηκεύει όλα τα διαθέσιμα δεδομένα και ταξινομεί ένα νέο σημείο δεδομένων με βάση την ομοιότητα. Αυτό σημαίνει ότι όταν εμφανίζονται νέα δεδομένα τότε μπορούν εύκολα να ταξινομηθούν σε μια κατηγορία χρησιμοποιώντας τον αλγόριθμο K-NN.
- Είναι ένας μη παραμετρικός αλγόριθμος, που σημαίνει ότι δεν κάνει καμία παραδοχή για τα υποκείμενα δεδομένα.
- Λέγεται αλγόριθμος «μαθητευόμενος τεμπέλης» επειδή δεν μαθαίνει από το σετ εκπαίδευσης αμέσως αλλά αποθηκεύει το σύνολο δεδομένων και κατά τη στιγμή της ταξινόμησης, εκτελεί μια ενέργεια στο σύνολο δεδομένων.
- Στη φάση της εκπαίδευσης αποθηκεύει το σύνολο δεδομένων και όταν λαμβάνει νέα δεδομένα, τότε ταξινομεί αυτά τα δεδομένα σε μια κατηγορία που μοιάζει πολύ με τα νέα δεδομένα.

[52] Alpydin (2010)
[53] Steinbach και Pang-Ning (2009)

Πλεονεκτήματα

- Είναι πολύ απλός.
- Η εκπαίδευση είναι ασήμαντη.
- Λειτουργεί με οποιονδήποτε αριθμό τάξεων.
- Εύκολη προσθήκη περισσότερων δεδομένων.

Μειονεκτήματα

- Το κόστος υπολογισμού είναι υψηλό λόγω του υπολογισμού της απόστασης μεταξύ των σημείων δεδομένων για όλα τα δείγματα εκπαίδευσης.
- Δεν είναι καλός με δεδομένα υψηλών διαστάσεων.

- SVM (Support Vectors Machines)[54]

Η μηχανή διανύσματος υποστήριξης SVM είναι μια μηχανή εκμάθησης για προβλήματα ταξινόμησης δύο ομάδων. Εφαρμόζει την ακόλουθη ιδέα: τα διανύσματα εισόδου χαρτογραφούνται μη γραμμικά σε έναν χώρο χαρακτηριστικών πολύ υψηλής διάστασης. Σε αυτό το χώρο χαρακτηριστικών κατασκευάζεται μια γραμμική επιφάνεια αποφάσεων. Οι ειδικές ιδιότητες της επιφάνειας αποφάσεων διασφαλίζουν υψηλή ικανότητα γενίκευσης της μηχανής εκμάθησης. Η ιδέα πίσω από το δίκτυο φορέα υποστήριξης εφαρμόστηκε προηγουμένως για την περιορισμένη περίπτωση όπου τα δεδομένα εκπαίδευσης μπορούν να διαχωριστούν χωρίς σφάλματα. Εδώ επεκτείνουμε αυτό το αποτέλεσμα σε μη διαχωρίσιμα δεδομένα εκπαίδευσης.

Αποδεικνύεται η υψηλή γενίκευση των δικτύων υποστήριξης-φορέων που χρησιμοποιούν μετασχηματισμούς πολυωνύμων εισόδων. Συγκρίνουμε επίσης την απόδοση του δικτύου υποστήριξης-φορέα με διάφορους αλγορίθμους κλασικής μάθησης που συμμετείχαν σε μια μελέτη αναφοράς της Αναγνώρισης Οπτικών Χαρακτήρων.

Τα πλεονεκτήματα της μεθόδου SVM είναι:

[54] Xue και λοιποί (2009)

- Αποτελεσματική σε χώρους υψηλών διαστάσεων.
- Αποτελεσματική ακόμα και σε περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων.
- Χρησιμοποιεί ένα υποσύνολο σημείων εκπαίδευσης στη συνάρτηση απόφασης (που ονομάζεται διάνυσμα υποστήριξης), επομένως είναι αποδοτικό και για τη μνήμη.
- Ευπροσάρμοστο: μπορούν να καθοριστούν διαφορετικές λειτουργίες πυρήνα για τη συνάρτηση απόφασης. Παρέχονται κοινοί πυρήνες, αλλά είναι επίσης δυνατό να καθοριστούν προσαρμοσμένοι πυρήνες.

Τα μειονεκτήματά της είναι:

- Δεν παρέχει άμεσα εκτιμήσεις πιθανότητας. Χρησιμοποιεί five-fold cross-validation.
- Δεν είναι κατάλληλοι για μεγάλα σύνολα δεδομένων λόγω του υψηλού χρόνου προπόνησης.

4.2 **Στόχος και σκοπός**

Με τη δημιουργία τεράστιων βάσεων δεδομένων και την ανάγκη για καλές τεχνικές μηχανικής μάθησης, προκύπτουν νέα προβλήματα και απαιτούνται νέες προσεγγίσεις για την επιλογή χαρακτηριστικών (F.S). Η επιλογή χαρακτηριστικών παίζει σημαντικό ρόλο στην ταξινόμηση και είναι ένα σημαντικό βήμα προεπεξεργασίας της μηχανικής μάθησης. Επιλέγει ένα αποτελεσματικό σύνολο από τα αρχικά χαρακτηριστικά σύμφωνα με ένα συγκεκριμένο κριτήριο, ώστε να μπορεί να βελτιώσει την απόδοση της μεταγενέστερης επεξεργασίας δεδομένων, όπως ταξινόμηση και ομαδοποίηση.

Στόχος μας είναι να χρησιμοποιήσουμε την Εντροπία ώστε να κάνουμε Επιλογή Χαρακτηριστικών (Feature Selection-FS) με απώτερο σκοπό να βελτιώσουμε ή να εκτιμήσουμε κατά πόσο συμβάλει η συγκεκριμένη διαδικασία FS στη βελτίωση των επερχόμενων εξατομικευμένων μοντέλων ταξινόμησης που θα αναπτύξουμε χρησιμοποιώντας μεθόδους υπολογιστικής νοημοσύνης και τα δεδομένα που περιγράφονται παρακάτω.

5 Πειραματική Συγκριτική Ανάλυση

5.1 Εισαγωγικά

Σκοπός αυτού του κεφαλαίου αλλά και της παρούσας εργασίας είναι να αξιολογήσει και να συγκρίνει την απόδοση μεθόδων εντροπίας σε προβλήματα απόφασης (Feature Selection), χρησιμοποιώντας μεθόδους ταξινόμησης.
Η ανάλυση αυτή θα βοηθήσει στον εντοπισμό αλληλεπιδράσεων που μπορεί να υπάρχουν μεταξύ της αποτελεσματικότητας των μεθόδων μέτρησης εντροπίας και των υποδειγμάτων ταξινόμησης, όπως και του εξεταζόμενου dataset.
Η έννοια της αποτελεσματικότητας αναφέρεται στη δυνατότητα εντοπισμού των χρήσιμων χαρακτηριστικών και στο ποσοστό μείωσης της εξεταζόμενης πληροφορίας.

Για την εξαγωγή των αποτελεσμάτων της συγκεκριμένης εργασίας χρησιμοποιήθηκε το πρόγραμμα Jupyter notebook (Anaconda 3) το οποίο χρησιμοποιεί τη γλώσσα Python.

Ο καρκίνος του μαστού αναφέρεται στην ανάπτυξη κακοήθους όγκου στην περιοχή του μαστού. Αποτελεί μια από τις συχνότερες εμφανιζόμενες μορφές καρκίνου παγκοσμίως και είναι η πρώτη σε αριθμό κρουσμάτων στον γυναικείο πληθυσμό. Προκαλείται από ανεξέλεγκτο πολλαπλασιασμό παθολογικών κυττάρων που ως αποτέλεσμα προκαλούν το σχηματισμό κακοήθους όγκου στην περιοχή του μαστού και ουσιαστικά αποτελεί κυτταρική νόσο. Τα παθολογικά αυτά κύτταρα έχουν τη δυνατότητα εξάπλωσης σε γειτονικούς ιστούς.

5.2 Παρουσίαση δεδομένων

Το σύνολο δεδομένων με τίτλο “Wisconsin Diagnostic Breast Cancer (WDBC)” είναι διαθέσιμο στο **http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)**. Περιέχει χαρακτηριστικά τα οποία απεικονίζουν τιμές περιοχών του στήθους γυναικών που

εξετάσθηκαν με σκοπό την διάγνωση καρκίνου του μαστού. Αυτό το έργο εξελίχθηκε μέσα από την επιθυμία του Dr. Wolberg να εντοπίσει με ακρίβεια τις μάζες του μαστού βασιζόμενος αποκλειστικά στη μέθοδο Fine Nose Vacuration (FNA). Κατάφερε να εντοπίσει εννέα οπτικά εκτιμημένα χαρακτηριστικά ενός δείγματος FNA, το οποίο θεωρούσε χρήσιμο για τη διάγνωση. Σε συνεργασία με τον καθηγητή Mangasarian και δύο από τους μεταπτυχιακούς φοιτητές του, Rudy Setiono και Kristin Bennett, κατασκευάστηκε ένας ταξινομητής χρησιμοποιώντας τη μέθοδο πολλαπλών επιφανειών (MSM) για τον διαχωρισμό προτύπων σε αυτά τα εννέα χαρακτηριστικά που διέγνωσε με επιτυχία το 97% των νέων περιπτώσεων. Το σύνολο δεδομένων που προκύπτει είναι γνωστό ως τα δεδομένα του καρκίνου του μαστού στο Wisconsin. Το έργο ανάλυσης εικόνας ξεκίνησε το 1990 με την προσθήκη του Nick Street στην ερευνητική ομάδα. Ο στόχος ήταν η διάγνωση του δείγματος με βάση μια ψηφιακή εικόνα ενός μικρού τμήματος της διαφάνειας FNA. Τα αποτελέσματα αυτής της έρευνας έχουν ενοποιηθεί σε ένα σύστημα λογισμικού γνωστό ως Xcyt, το οποίο χρησιμοποιείται σήμερα από τον Dr.Wolberg στην κλινική πρακτική του. Η διαδικασία διάγνωσης εκτελείται τώρα ως εξής: Ένα FNA δείγμα λαμβάνεται από τη μάζα του μαστού. Αυτό το υλικό στη συνέχεια τοποθετείται σε μια πλάκα μικροσκοπίου και χρωματίζεται για να επισημάνει τους κυτταρικούς πυρήνες. Ένα τμήμα της ολίσθησης στην οποία τα κύτταρα είναι καλά διαφοροποιημένα σαρώνονται στη συνέχεια χρησιμοποιώντας μια ψηφιακή φωτογραφική μηχανή και ένα πίνακα πλαισίου. Ο χρήστης στη συνέχεια απομονώνει τους μεμονωμένους πυρήνες χρησιμοποιώντας Xcyt. Χρησιμοποιώντας τον δείκτη του ποντικιού, ο χρήστης σχεδιάζει το κατά προσέγγιση όριο κάθε πυρήνα. Στη συνέχεια προσεγγίζοντας το πρόβλημα μέσα από μια οπτική των υπολογιστών, γνωστή και ως "snakes", οι προσεγγίσεις αυτές συγκλίνουν στη συνέχεια στα ακριβή πυρηνικά όρια. Αυτή η διαδραστική διαδικασία διαρκεί από δύο έως πέντε λεπτά ανά διαφάνεια. Μόλις απομονωθούν όλοι (ή οι περισσότεροι) από τους πυρήνες σε αυτό το φάσμα, το πρόγραμμα υπολογίζει τιμές για καθένα από τα δέκα χαρακτηριστικά κάθε πυρήνα, μετρώντας το μέγεθος, το σχήμα και την υφή.

	Χαρακτηριστικά	**Εύρος τιμών**
1	Sample Code	ID Number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoni	1-10
10	Mitoses	1-10
11	Class	2 (Benign) ή 4 (Malignant)

Πληροφορίες χαρακτηριστικών:

1) ID number

2) Diagnosis (M = malignant, B = benign)

Για κάθε πυρήνα κυττάρων υπολογίζονται δέκα χαρακτηριστικά:

1) radius (μέσος όρος αποστάσεων από το κέντρο στα σημεία της περιμέτρου)

2) texture (τυπική απόκλιση τιμών γκρι κλίμακας)

3) perimeter

4) area

5) smoothness (τοπική μεταβολή στα μήκη ακτίνας)

6) compactness (perimeter^2 / area - 1.0)

7) concavity (βαρύτητα κοίλων τμημάτων του περιγράμματος)

8) concave points (αριθμός κοίλων τμημάτων του περιγράμματος)

9) symmetry

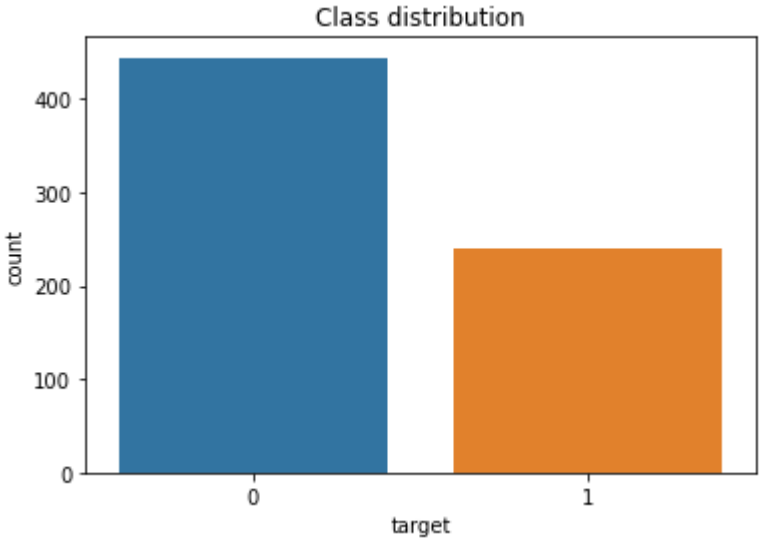
10) fractal dimension ("coastline approximation" - 1)

Το dataset αποτελείται από 10 στήλες στο σύνολο και 683 γραμμές. Αφαιρέσαμε 16 γραμμές διότι τα δεδομένα ήταν ελλιπή. Στόχος μας είναι η εύρεση του αλγορίθμου μηχανικής μάθησης που θα έχει σαν αποτέλεσμα το μεγαλύτερο ποσοστό ακρίβειας στην εκτίμηση των περιπτώσεων καρκίνου του μαστού, σε καλοήθεις (αρνητικές σε καρκίνο) και κακοήθεις (θετικές σε καρκίνο).

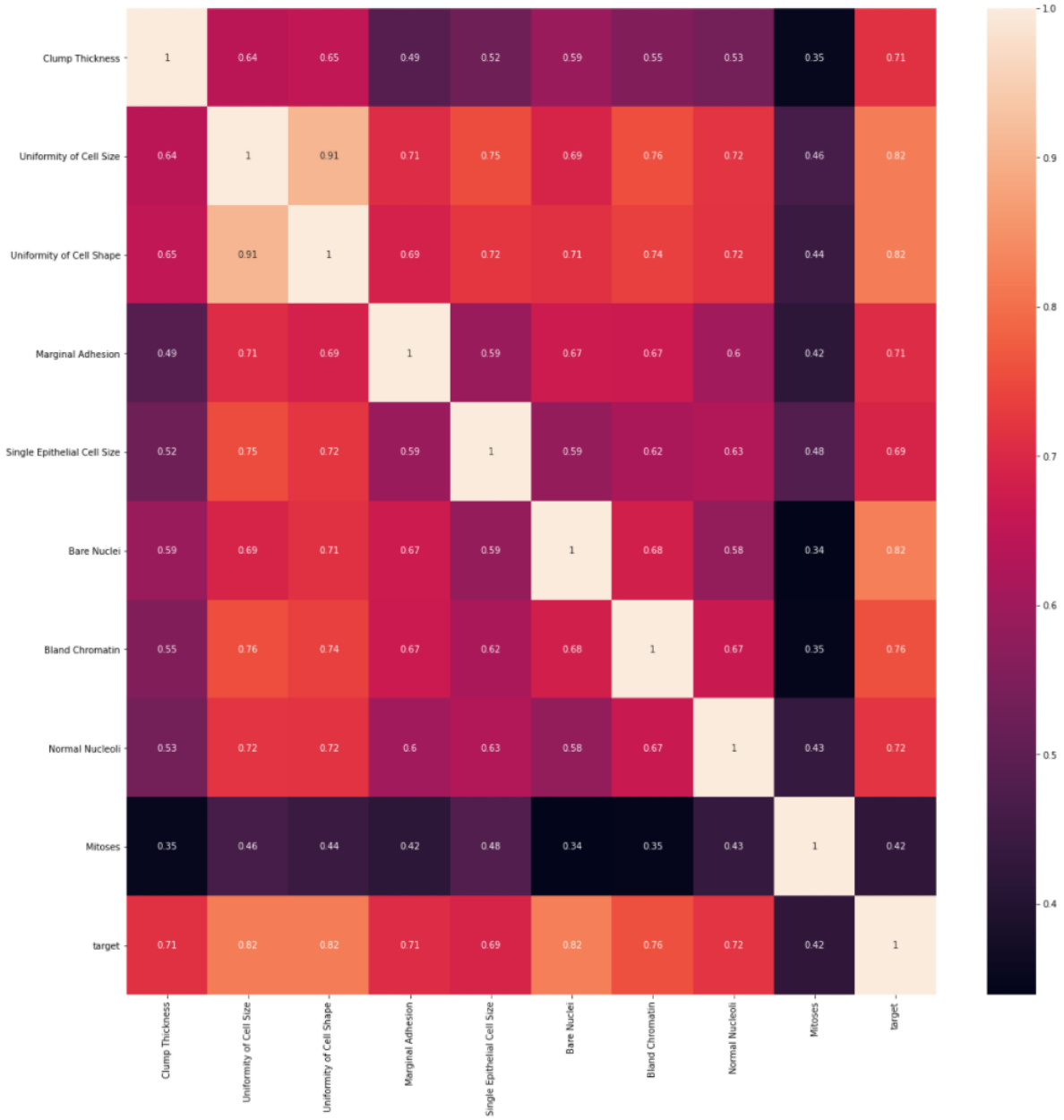
5.3 Εισαγωγή και προετοιμασία των δεδομένων

Αρχικά, εισάγουμε τα δεδομένα (WBCDS) στο πρόγραμμα. Η στήλη Class, που αποτελεί την κλάση του dataset με τιμές M (Malignant) και B (Benign), βλέπουμε ότι δεν αποτελείται από συνεχείς τιμές κι έτσι αντικαθιστούμε τις τιμές αυτές με 0 και 1 (M:1, B:0), ώστε να βοηθήσει την ανάλυσή μας. Αλλάζουμε τη στήλη Class σε Target και τέλος, αφαιρούμε τελείως τη στήλη Sample Code διότι δεν προσφέρει κάποια πληροφορία. Επομένως, καταλήγουμε σε ένα σύνολο δεδομένων με 683 παρατηρήσεις (458 καλοήθεις, 241 κακοήθεις), 10 χαρακτηριστικά και δύο κλάσεις.

Στον παρακάτω πίνακα βλέπουμε την κατανομή των παρατηρήσεων μας ανά κλάση. Οι παρατηρήσεις με κλάση 0 είναι στο σύνολο 458 και με κλάση 1 είναι 241. Η κατανομή μας μοιάζει φυσιολογική.



Στο παρακάτω σχήμα διαπιστώνουμε πως υπάρχει πολύ μεγάλη συσχέτιση μεταξύ ορισμένων χαρακτηριστικών. Συγκεκριμένα, το Uniformity of Cell Size σχετίζεται πάρα πολύ με το Uniformity of Cell Shape (0.91). Στη συνέχεια, για να αποφύγουμε το overfitting λόγω αλληλοσχετιζόμενων χαρακτηριστικών θα εφαρμόσουμε συγκεκριμένες μεθόδους μέτρησης εντροπίας ώστε να κάνουμε επιλογής χαρακτηριστικών (F.S).



6 Παρουσίαση αποτελεσμάτων μεθόδων εντροπίας

Στην συγκεκριμένη ενότητα παρουσιάζονται τα αποτελέσματα των μεθόδων εντροπίας mutual information, Shannon Entropy και Renyi Entropy που εφαρμόστηκαν πάνω στο dataset Wisconsin breast cancer. Συγκεκριμένα καταγράφονται τα σύνολα δεδομένων όπως αυτά διαμορφώθηκαν μετά την εφαρμογή της κάθε μεθόδου, όπως επίσης οι επιδόσεις αλγορίθμων ταξινόμησης τόσο στα συγκεκριμένα σύνολα όσο και στο αρχικό dataset. Οι αλγόριθμοι ταξινόμησης που τρέξαμε είναι ο SVM και ο Logistic Regression

6.1 **Mutual Information**

Στον παρακάτω πίνακα αποτυπώνονται, οι τιμές πληροφορίας των χαρακτηριστικών του συνόλου δεδομένων εφαρμόζοντας τη μέθοδο mutual information καθώς και οι τιμές εντροπίας εφαρμόζοντας τη μέθοδο Shannon Entropy. Χρησιμοποιώντας τη μεθόδους αυτές θα επιλέξουμε τα χαρακτηριστικά με τις υψηλότερες τιμές.

Χαρακτηριστικά	Mutual Information	Shannon Entropy
Uniformity of Cell Size	0.49893957	2.34
Uniformity of Cell Shape	0.47819086	2.49
Bare Nuclei	0.42240375	2
Single Epithelial Cell Size	0.38415377	2.29
Bland Chromatin	0.37938646	2.77
Clump Thickness	0.33332967	3.05
Normal Nucleoni	0.32181526	2.05
Marginal Adhesion	0.29709053	2.21
Mitoses	0.124162	1.13

Πίνακας 2 Τιμές εντροπίας με mutual information και Shannon entropy.

7 Αποτελέσματα ταξινόμησης

7.1 **Αποτελέσματα SVM**

Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγόριθμου ταξινόμησης SVM στο αρχικό dataset καθώς και στα υποσύνολα που δημιουργήθηκαν έπειτα από την εφαρμογή συγκεκριμένων μεθόδων εντροπίας.

Data Set	Accuracy	Precision	Recall	f 1-score
SVM All	94.9	91.8	93.7	92.8
SVM mi	94.1	91.6	91.6	91.6
SVM (SE)	93.4	91.5	89.5	90.5
SVM (Re)	93.4	91.5	89.6	90.5

Πίνακας 3 Αποτελέσματα ταξινομητή SVM

Data Set	Accuracy	Precision	Recall	f 1-score
SVM All	96.6	95.1	95.4	95.2
SVM mi	96.2	95.6	93.7	94.4
SVM (SE)	96.5	94.3	95.8	95
SVM (Re)	96	93.6	95.4	94.4

Πίνακας 4 Αποτελέσματα ταξινομητή SVM εφαρμόζοντας 10-fold cross validation

Αναλύοντας τους παραπάνω πίνακες, διαπιστώθηκε πως ο ταξινομητής SVM επιτυγχάνει υψηλό ποσοστό επιτυχίας, το οποίο αγγίζει το 95%. Επίσης, παρατηρείται πως το ποσοστό επιτυχίας του ταξινομητή, ανάλογα με την εκάστοτε μέθοδο FS που επιλέχθηκε, παρουσιάζει μικρές διαφορές. Συγκεκριμένα, η μέθοδος Mutual Information παρουσιάζει μεγαλύτερη ακρίβεια με ποσοστό επιτυχίας 94.1% με την μέθοδο Shannon Entropy να ακολουθεί με 93.4%. Εφαρμόζοντας την πολλαπλή επικύρωση ten fold cross validation, παρατηρείται πως το ποσοστό επιτυχίας του ταξινομητή παραμένει σε υψηλά επίπεδα, με την μέθοδο Shannon Entropy αυτή τη φορά να παρουσιάζει το υψηλότερο ποσοστό ακριβείας με 96.5%.

Θα πρέπει επίσης να εστιάσουμε την ανάλυση μας και στο μέτρο Recall, το οποίο ουσιαστικά αποτυπώνει πόσες από τις πραγματικά θετικές παρατηρήσεις προβλέπονται σωστά. Το συγκεκριμένο μέτρο ίσως είναι και το σημαντικότερο απ' όλα καθώς η σωστή πρόβλεψη των πραγματικά θετικών παρατηρήσεων είναι καθοριστική. Σύμφωνα με το μέτρο Recall και την πολλαπλή επικύρωση ten fold cross validation, η μέθοδος Shannon Entropy είναι αποτελεσματικότερη από την Mutual Information.

7.2 **Αποτελέσματα Logistic Regression**

Όπως και παραπάνω, εφαρμόσαμε τον αλγόριθμο ταξινόμησης Logistic Regression στο αρχικό dataset καθώς και στο διαμορφωμένο dataset μετά την επιλογή χαρακτηριστικών που πραγματοποιήσαμε.

Data Set	Accuracy	Precision	Recall	f 1-score
LR All	95.6	93.7	93.7	93.7
LR mi	94.3	91.8	91.7	91.8
LR (SE)	93.4	91.5	89.6	90.5
LR (Re)	93.4	91.5	89.6	90.5

Πίνακας 5 Αποτελέσματα ταξινομητή Logistic Regression

Data Set	Accuracy	Precision	Recall	f 1-score
LR All	96.6	95.1	95.4	95.2
LR mi	96.2	95.6	93.7	94.4
LR (SE)	96.4	94.4	95.8	95
LR (Re)	96	93.6	95.4	94.4

Πίνακας 6 Αποτελέσματα ταξινομητή Logistic Regression εφαρμόζοντας 10-fold cross validation

Βλέποντας τα αποτελέσματα στους πίνακες φαίνεται πως και ο ταξινομητής Logistic Regression επιτυγχάνει υψηλά ποσοστά επιτυχίας. Παρατηρούνται και πάλι πολύ μικρές διαφορές στο ποσοστό επιτυχίας του ταξινομητή, ανάλογα την μέθοδο FS που επιλέχθηκε.

Ειδικότερα, τη μεγαλύτερη ακρίβεια παρουσιάζει και πάλι η μέθοδος Mutual Information, 95.6%, με την μέθοδο Shannon Entropy και Renyi Entropy να ακολουθούν με το ίδιο ποσοστό ακριβείας. Εφαρμόζοντας πολλαπλή επικύρωση ten fold cross validation και παρατηρώντας παράλληλα τον δείκτη Recall, ο οποίος αποτελεί ίσως το σημαντικότερο κριτήριο, διαπιστώνουμε πως η μέθοδος Shannon Entropy επιτυγχάνει με πολύ μικρή διαφορά το μεγαλύτερο ποσοστό επιτυχίας. Ακολουθούν σε φθίνουσα σειρά η μέθοδος Renyi Entropy και Mutual Information.

8 Επίλογος

8.1 **Σύνοψη**

Η σύγκριση των ποσοστών ακριβείας για τους δύο ταξινομητές (SVM, Logistic Regression) έδειξε ότι με ή χωρίς την εφαρμογή επιλογής χαρακτηριστικών στο σύνολο δεδομένων WDBC η ακρίβεια του ταξινομητή SVM είναι ίση με την ακρίβεια του ταξινομητή Logistic Regression. Επίσης, παρατηρείται ότι η επιλογή χαρακτηριστικών δεν βελτίωσε την ακρίβεια των ταξινομητών. Εφαρμόζοντας πολλαπλή επικύρωση ten fold cross validation και μελετώντας το μέτρο Recall παρατηρείται πως η επιλογή χαρακτηριστικών, με τη μέθοδο Shannon Entropy, βελτίωσε την απόδοση και των δύο ταξινομητών (95.8%) ενώ η μέθοδος Renyi Entropy κράτησε την απόδοση τους στα ίδια επίπεδα (95.4%).

8.2 **Συζήτηση και προτάσεις**

Τα τελευταία χρόνια, η εφαρμογή μεθόδων επιλογής χαρακτηριστικών σε ιατρικά σύνολα δεδομένων έχει αυξηθεί πολύ. Το δύσκολο έργο στην επιλογή χαρακτηριστικών είναι πώς να αποκτήσουμε ένα βέλτιστο υποσύνολο σχετικών και μη περιττών χαρακτηριστικών που θα δώσει μια βέλτιστη λύση χωρίς να αυξήσει την πολυπλοκότητα της εργασίας μοντελοποίησης. Επομένως, υπάρχει ανάγκη να ενημερωθούν οι επαγγελματίες για τις

μεθόδους επιλογής χαρακτηριστικών που έχουν εφαρμοστεί με επιτυχία σε σύνολα ιατρικών δεδομένων και να επισημανθούν οι μελλοντικές τάσεις σε αυτόν τον τομέα. Η ανάπτυξη μιας καθολικής μεθόδου που επιτυγχάνει την καλύτερη ακρίβεια ταξινόμησης με λιγότερα χαρακτηριστικά εξακολουθεί να είναι ένας ανοιχτός τομέας έρευνας.

Βιβλιογραφία

[1] Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. Newberry Park.

[2] Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society: Series B (Methodological), 28(1), 131-142.

[3] Alpaydin, E. (2010). Design and analysis of machine learning experiments.

[4] Arimoto, S. (1971). Information-theoretical considerations on estimation problems. Information and control, 19(3), 181-194.

[5] Arndt, C. (2003). Information measures: information and its description in science and engineering. Springer Science & Business Media.

[6] Barnes, D. J., Fincher, S., & Thompson, S. (1997). Introductory problem solving in computer science. In 5th Annual Conference on the Teaching of Computing (pp. 36-39).

[7] Bhandari, D., & Pal, N. R. (1993). Some new information measures for fuzzy sets. Information Sciences, 67(3), 209-228.

[8] Burbea, J. (1984). The Bose-Einstein Entropy Of Degree-Alpha And Its Jensen Difference. Utilitas Mathematica, 25(MAY), 225-240.

[9] Cover, T. M. (1999). Elements of information theory. John Wiley & Sons.

[10] De Luca, A., & Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and control, 20(4), 301-312.

[11] Edwards, S. (2008). Elements of Information Theory, Thomas M. Cover, Joy A. Thomas, John Wiley & Sons, Inc.(2006).

[12] Favretti, M. (2018). Remarks on the maximum entropy principle with application to the maximum entropy theory of ecology. Entropy, 20(1), 11.

[13] Ferreri, C. (1980). Hypoentropy and related heterogeneity, divergency and information measures.

[14] Gallager, R. G. (1968). Information theory and reliable communication (Vol. 2). New York: Wiley.

[15] Havrda, J., & Charvát, F. (1967). Quantification method of classification processes. Concept of structural $ a $-entropy. Kybernetika, 3(1), 30-35.

[16] IGNOU (2017) Unit-7 Information Theory: Measure and Contents Evaluation.

[17] Kapur, J. N. (1989). Maximum-entropy models in science and engineering. John Wiley & Sons.

[18] Kesavan, H. K. (2009). Jaynes' Maximum Entropy Principle.

[19] Kosko, B. (1986). Fuzzy entropy and conditioning. Information sciences, 40(2), 165-174.

[20] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86.

[21] Learned-Miller, E. G. (2013). Entropy and mutual information. Department of Computer Science, University of Massachusetts, Amherst.

[22] Lee, W., & Xiang, D. (2000, May). Information-theoretic measures for anomaly detection. In Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001 (pp. 130-143). IEEE.

[23] Li, X., & Liu, B. (2006). Cross-entropy and generalized entropy for fuzzy variables. Technical Report.

[24] Liese, F., & Vajda, I. (2006). On divergences and informations in statistics and information theory. IEEE Transactions on Information Theory, 52(10), 4394-4412.

[25] Lombardi, O., Holik, F., & Vanni, L. (2016). What is Shannon information?. Synthese, 193(7), 1983-2012.

[26] Markowitz, H. M. (1959). Portfolio Selection: E cient Diversification of Investments. Cowles Foundation Monograph, 16.

[27] Maszczyk, T., & Duch, W. (2008). Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. In International Conference on Artificial Intelligence and Soft Computing (pp. 643-651). Springer, Berlin, Heidelberg.

[28] Michalowicz, J. V., Nichols, J. M., & Bucholtz, F. (2008). Calculation of differential entropy for a mixed Gaussian distribution. Entropy, 10(3), 200-206.

[29] Nychis, G., Sekar, V., Andersen, D. G., Kim, H., & Zhang, H. (2008, October). An empirical evaluation of entropy-based traffic anomaly detection. In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (pp. 151-156).

[30] Omran, M. G., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. Intelligent Data Analysis, 11(6), 583-605.

[31] Ou, J. (2005). Theory of portfolio and risk based on incremental entropy. The Journal of Risk Finance.

[32] Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, 3(3), 370-379.

[33] Papaleo, L. (2014). INTRODUCTION TO XML AND ITS APPLICATIONS. In Handbook of Metadata, Semantics and Ontologies (pp. 109-139).

[34] Pincus, S. M., Gladstone, I. M., & Ehrenkranz, R. A. (1991). A regularity statistic for medical data analysis. Journal of clinical monitoring, 7(4), 335-345.

[35] Prehl, J., Essex, C., & Hoffmann, K. H. (2012). Tsallis relative entropy and anomalous diffusion. Entropy, 14(4), 701-716.

[36] Rényi, A. (1961). On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (pp. 547-561). University of California Press.

[37] Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

[38] Sharma, B. D., & Mittal, D. P. (1975). New non-additive measures of entropy for discrete probability distributions. J. Math. Sci, 10, 28-40.

[39] Steinbach, M., & Tan, P. N. (2009). kNN: k-nearest neighbors. In The top ten algorithms in data mining (pp. 165-176). Chapman and Hall/CRC.

[40] Stone, J. V. (2015). Information theory: a tutorial introduction.

[41] Thomas, J. A., & Cover, T. M. (1991). Elements of information Theory. John Willey & Sons. Inc, New York.

[42] Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. Journal of statistical physics, 52(1), 479-487.

[43] Vajapeyam, S. (2014). Understanding shannon's entropy metric for information. arXiv preprint arXiv:1405.2061.

[44] Varma, R. S. (1966). Generalizations of Renyi's entropy of order α. Journal of Mathematical Sciences, 1(7), 34-48.

[45] Xue, H., Yang, Q., & Chen, S. (2009). SVM: Support vector machines. In The top ten algorithms in data mining (pp. 51-74). Chapman and Hall/CRC.

[46] Zenil, H. (2020). Towards Demystifying Shannon Entropy, Lossless Compression and Approaches to Statistical Machine Learning. In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 47, No. 1, p. 24).

[47] Ζορκάδης, Βασίλειος. «Θεωρία Πληροφορίας και Κωδικοποίησης». Βασικά ζητήματα δικτύων Η/Υ, Τόμος Α΄, ΕΑΠ.

[48] Πούλος, Μ. 2015. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ: Εισαγωγή. [Κεφάλαιο Συγγράμματος]. Στο Πούλος, Μ. 2015. Σημασιολογική επεξεργασία της πληροφορίας. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 1. Διαθέσιμο στο: http://hdl.handle.net/11419/2855.

[49] Χάλκος, Γ. 2011 «Οικονομετρία – Θεωρία, εφαρμογές & χρήση προγραμμάτων σε Η/Υ», Gutenberg, Αθήνα.