Τ.Ε.Ι ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ: ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΘΕΜΑ:
"Τεχνικές Ανάλυσης Μεγάλων Δεδομένων"

Πτυχιακή Εργασία: Μεχίλι Μαρία
Χριστοπούλου Ευσταθία

Επιβλέπων Καθηγητής Γεράσιμος Αντζουλάτος

ΠΑΤΡΑ 2015

Περίληψη

Η μεγαλύτερη πρόκληση των σύγχρονων υπολογιστικών συστημάτων είναι αναμφισβήτητα η αποδοτική αποθήκευση και ανάκτηση πολύ μεγάλου όγκου δεδομένων. Η ανάγκη αυτή έκανε την εμφάνισή της τα τελευταία χρόνια λόγω της έκρηξης δεδομένων που παρατηρείται στο διαδίκτυο και αποκτά ολοένα και μεγαλύτερη σημασία λόγω του πολύ μεγάλου εύρους πληροφοριών που μπορούμε να αντλήσουμε. Στην παρούσα εργασία μελετάμε τα μεγάλης κλίμακας Δεδομένα και πως αυτά αναλύονται χρησιμοποιώντας συστήματα Επιχειρηματικής Εφυιας.Συγκεκριμενα αναλύουμε τη σπουδαιότητα των μεγάλων δεδομένων ,τις περιπτώσεις χρήσης τους, το πρότυπο των 3v. Διακρίνουμε το κατά πόσο είναι απαραίτητα στη λήψη σωστών, έγκυρων και έγκαιρων αποφάσεων ως παράγοντα επιτυχίας για τις περισσότερες σύγχρονες επιχειρήσεις και οργανισμούς. Ταυτόχρονα, τα τελευταία χρόνια, με την ανάπτυξη νέων τεχνολογιών και εφαρμογών – όπως η εξάπλωση των κοινωνικών δικτύων, η εκτεταμένη χρήση smart phones, η εγκατάσταση αισθητήρων κ.α. – ο όγκος και η μορφή των δεδομένων έχει αλλάξει δραματικά, ενώ οι δυνατότητες ανάλυσης και επεξεργασίας αυτών είναι εντυπωσιακές .Έτσι προσπαθούμε να αναλύσουμε και τις τεχνολογίες μεγάλων δεδομένων ,τις εφαρμογές τους και τις τεχνικές ανάλυσης τους.

ABSTRACT

The biggest challenge of modern computing is undoubtedly the efficient storage and retrieval of very large amounts of data .This need made its appearance in recent years due to the explosion of data that is observed on the Internet and is becoming more important because of the very wide range of information we can learn. In the present study we study the large-scale data and how they are analyzed using business intelligence systems. Specifically we analyze the importance of large data, the cases of use, the model of 3V.We distinguish whether it is necessary to take proper, accurate and timely decisions as success factor for most modern businesses and organizations. At the same time the last few years with the development of new technologies and applications - such as the spread of social networks, the widespread use of smart phones, the installation of sensors etc. - The volume and the data format has changed dramatic, and analytical capacity and processing them is impressive. So we try to analyze big data technologies, their applications and their analysis techniques.

Πίνακας περιεχομένων

ΕΙΣΑΓΩΓΗ

1.1ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Την Δεκαετία 1950 για την διαχείριση και επεξεργασία μεμονωμένων αρχείων χρησιμοποιούνταν οι κάρτες και οι ταινίες. Οι εξελίξεις σε συσκευές μαζικής αποθήκευσης τυχαίας πρόσβασης και αύξηση υπολογιστικής ισχύος θέτουν τις προϋποθέσεις για την ανάπτυξη συστημάτων διαχείρισης δεδομένων σε αντικατάσταση των συστημάτων διαχείρισης αρχείων.

Την Δεκαετία 1960 τα πρώτα συστήματα διαχείρισης βάσεων δεδομένων δημιουργήθηκαν τη δεκαετία του 1960 με σκοπό ένα κοινό οργανωτικό πλαίσιο για την διαχείριση δεδομένων τα οποία μέχρι τότε αποθηκεύονταν σε μεμονωμένα αρχεία. Το 1964, ο Charles Bachman της General Electric πρότεινε ένα δικτυωτό μοντέλο δεδομένων(network data model) στο οποίο οι εγγραφές δεδομένων ήταν συνδεδεμένες μεταξύ τους με τέτοιο τρόπο ώστε να σχηματίζουν τεμνόμενα σύνολα δεδομένων. Τα πρώτα συστήματα διαχείρισης βάσεως δεδομένων στηρίχθηκαν σε αυτό το δικτυωτό μοντέλο. Το 1965 η εταιρία IBM και η διεύθυνση διαστήματος της North American Aviation ανέπτυξαν από κοινού το ιεραρχικό μοντέλο δεδομένων. Σε αυτό το μοντέλο, τα δεδομένα παριστάνονταν ως δενδροειδής δομές μέσα σε μια ιεραρχία εγγράφων. Το Σύστημα Διαχείρισης Πληροφοριών(information management system-IMS) της IBM που κυκλοφόρησε το 1969 ήταν βασιζόμενο στο ιεραρχικό μοντέλο δεδομένων. Από τα δικτυωτά και ιεραρχικά συστήματα μόνο τα IMS παραμένει σε χρήση μέχρι και σήμερα.

Την Δεκαετία 1970 ο ορισμός του σχεσιακού μοντέλου δεδομένων έγινε για πρώτη φορά το 1970 από τον Edgar Codd σε ένα ερευνητικό έντυπο της IBM με τίτλο ‘System R4 Relational’. Στην αρχή βέβαια δεν ήταν ξεκάθαρο κατά πόσο ένα σχεσιακό σύστημα που θα βασιζόταν στο σχεσιακό μοντέλο θα μπορούσε να πετύχει εμπορικά. Έτσι μέχρι και το 1979 όλες οι εμπορικές υλοποιήσεις βάσεων δεδομένων βασίζονταν είτε στην δικτυωτή είτε στην ιεραρχική προσέγγιση. Άρχισαν βέβαια να αναπτύσσονται βέβαια τα ερευνητικά προγράμματα σχεσιακών συστημάτων System R(IBM) και INGRESS καθώς και σχεσιακές γλώσσες SEQUEL, QBE και QUEL. Το

1976 το μοντέλο οντοτήτων-σχέσεων(ER-Entity Relationship model) προτάθηκε από τον P.P. CHEN για να περιγράψει με γραφικά σύμβολα τα δεδομένα ως οντότητες, συσχετίσεις(σχέσεις) και γνωρίσματα. Το 1979 ιδρύθηκε η εταιρία Relational Software Incorprorated και κυκλοφόρησε στην αγορά την σχεσιακή βάση δεδομένων ORACLE V.2.

Την Δεκαετία 1980 η σχεσιακή γλώσσα SQL(μέρος του system R) αντικατέστησε την QUEL στο σύστημα INGRESS. Αναπτύχθηκαν οι έννοιες της διαχείρισης συναλλαγών(transaction management) από τον Jim Gray. Οι τάσεις που άρχιζαν να εμφανίζονται εκείνη την περίοδο αφορούσαν τα αντικειμενοστραφή συστήματα, την αρχιτεκτονική πελάτη-διακομιστή και τις κατανεμημένες βάσεις. Οι εγκαταστάσεις των σχεσιακών συστημάτων αυξάνουν με γοργούς ρυθμούς με πρώτα τα συστήματα Oracle, Server, SQL, Sybase, Informix, DB2. Εμφανίζονται τα σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων και σε προσωπικούς υπολογιστές : Dbase εξελίχθηκε μέχρι τις μέρες μας σε Paradox και η πιο γνωστή Microsoft Access.

Την Δεκαετία 1990 εμφανίζονται τα πρώτα εμπορικά αντικειμενοστραφή συστήματα Βάσεων Δεδομένων, η σύνδεση ΒΔ στο διαδίκτυο. Διαδίδεται ευρύτατα η τεχνολογία που επιτρέπει την επικοινωνία των χρηστών με ΒΔ μέσω διαδικτύου(HTML, ASP, XML).

Το 1991 το διαδίκτυο, ο παγκόσμιος ιστός όπως τον ξέρουμε, γεννιέται. Το πρωτόκολλο μεταφοράς υπερκειμένων (HTTP) γίνεται το βασικό μέσον διαμοιρασμού πληροφοριών.

Το 1995 η Sun βγάζει στην κυκλοφορία την πλατφόρμα Java. Η Java, που ανακαλύφθηκε το 1991, γίνεται η δεύτερη πιο διαδεδομένη γλώσσα μετά την C. Κυριαρχεί στις εφαρμογές διαδικτύου και καθιερώνεται στις μεσαίου επιπέδου εφαρμογές. Αυτές οι εφαρμογές είναι η πηγή καταγραφής και αποθήκευσης της κίνησης του διαδικτύου.

Το παγκόσμιο σύστημα εντοπισμού (GPS) γίνεται πλήρως λειτρουργικό. Το GPS είχε αναπτυχθεί αρχικά απο την DAPRA (υπηρεσία προγραμμάτων προηγμένης έρευνας και άμυνας) για στρατιωτικές εφαρμογές στις αρχές της δεκαετίας του ’70. Σήμερα η τεχνολογία αυτή είναι πανταχού παρούσα, απο εφαρμογές πλοήγησης αυτοκινήτων και αεροπλάνων μέχρι την ανεύρεση χαμένων τηλεφώνων iPhone.

Το 1998 ο Carlo Strozzi αναπτύσσει μια ανοιχτού κώδικα βάση δεδομένων και την αποκαλεί NoSQL. Δέκα χρόνια αργότερα, η πρωτοβουλία ανάπτυξης βάσεων δεδομένων NoSQL που θα μπορεί να επεξεργάζεται μεγάλα και αδόμητα σύνολα δεδομένων, κερδίζει ολοένα και περισσότερο έδαφος.Ιδρύεται η Google απο τους Larry Page και Sergey Brin οι οποίοι έχουν εργαστεί για περίπου έναν χρόνο σε ενα έργο μηχανής αναζήτησης του πανεπιστημίου Stanford με την ονομασία BackRub.

Το 1999 ο Kevin Ashton, συνιδρυτής του κέντρου Auto-ID στο ινστιτούτο τεχνολογίας της Μασσαχουσέτης (MIT) ανακαλύπτει τον όρο «Το διαδίκτυο των πραγμάτων»

Το 2001 ξεκινά η λειτουργία του Wikipedia. Μια εγκυκλοπαίδεια πληθώρας πηγών που φέρνει την επανάσταση στον τρόπο με τον οποίο οι άνθρωποι αναζητούν πληροφορίες.

Το 2002 το Ινστιτούτο Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών (IEEE) ορίζει την πρώτη έκδοση (1.1) των προδιαγραφών Bluetooth. Το Bluetooth ειναι μια ασύρματη τεχνολογία μεταφοράς δεδομένων σε μικρές αποστάσεις. Η εξέλιξη αυτών των προδιαγραφών και η υιοθέτησή τους οδήγησε σε μια νέα σειρά φορητών συσκευών και επέτρεψε την επικοινωνία μεταξύ της συσκευής αυτής και ενός άλλου υπολογιστή. Σήμερα σχεδόν κάθε φορητή συσκευή έχει και δέκτη Bluetooth.

Το 2003 σύμφωνα με μελέτες του IDC και EMC, ο όγκος δεδομένων που δημιουργήθηκε το 2003 ξεπερνά εκείνον που είχε δημιουργηθεί σε ολόκληρη την ιστορία της ανρθρωπότητας μέχρι τη στιγμή αυτή. Εκτιμάται οτι 1.8 zettabytes δεδομένων δημιουργήθηκαν μόνο το 2011. (1.8 zettabytes ισοδυναμούν με 200 δισεκατομμύρια ταινίες υψηλής ευκρίνειας HD, διάρκειας 2 ωρών έκαστος, ή αλλιώς με προβολή οπτικοακουστικού υλικού διάρκειας 47 εκατομμυρίων ετών χωρίς διακοπή για τουαλέττα.) Ξεκινά η λειτουργία του LinkedIn, του δημοφιλούς μέσου κοινωνικής διαδικτύωσης για επαγγελματίες. Το 2013 η ιστοσελίδα είχε περίπου 260 εκατομμύρια χρήστες.

Το 2004 τον Φεβρουάριο, το Wikipedia φθάνει τα 500.000 άρθρα και επτά μήνες αργότερα ξεπερνά το 1 εκατομμύριο.Η υπηρεσία κοινωνικής δικτύωσης Facebook ιδρύεται απο τον Mark Zuckerberg και άλλους στο Cambridge της Μασσαχουσέτης. Το 2013, η ιστοσελίδα έχει πάνω απο 1.15 δισεκατομμύρια χρήστες.

Το 2005 το ερευνητικό έργο Apache Hadoop δημιουργείται απο τους Doug Cutting και Mike Caferella. Το όνομα του έργου προήλθε απο το παιδικό παιχνίδι ενός ελέφαντα του γιού του Cutting. Ο σήμερα πλέον, διάσημος κίτρινος ελέφαντας, έγινε οικεία λέξη μόλις λίγα χρόνια αργότερα και θεμελιώδες μέρος σχεδόν όλων των στρατηγικών μεγάλων δεδομένων.Το Εθνικό Επιστημονικό Συμβούλιο προτείνει στο Εθνικό Ιδρυμα Επιστημών την δημιουργία μιας επαγγελματικής οδού για «εναν επαρκή αριθμό υψηλής ποιότητας επιστημόνων» που θα διαχειριστούν την αυξανόμενη συλλογή των ψηφιακών πληροφοριών.

Το 2007 η Apple βγάζει στην κυκλοφορία το iPhone και δημιουργεί μια ισχυρή καταναλωτική αγορά για έξυπνα κινητά τηλέφωνα.

Το 2008 ο αριθμός των συσκευών που είναι συνδεδεμένα στο διαδίκτυο ξεπερνά τον παγκόσμιο πληθυσμό.

Το 2011 ο υπολογιστής Watson της IBM σαρώνει και αναλύει 4 terabytes (200 εκατομμύρια σελίδες) δεδομένων σε δευτερόλεπτα νικώντας δύο ανθρώπινους παίκτες στο τηλεοπτικό πρόγραμμα “Jeopardy”! (περισσότερα για το τηλεοπτικό πρόγραμμα στο δέυτερο μέρος).Ξεκινά εργασία σε UnQL , μια γλώσσα επερωτήσεων για βάσεις δεδομένων NoSQL.Τα διαθέσιμα στοιχεία στην διεύθυνση του IPv4 έχουν όλα χρησιμοποιηθεί. Το IPv4 είναι μια σταθερά για να ορίζεται η διεύθυνση του πρωτοκόλλου διαδικτύου (IP). Το πρωτόκολλο IPv4 βασίστηκε σε εναν 32μπιτο αριθμό που σημαίνει οτι υπάρχουν διαθέσιμες 2^32 ή 4.5 δισεκατομμύρια μοναδικές διευθύνσεις. Το γεγονός αυτό μαρτυρά την πραγματική ζήτηση και ποσότητα των συνδεδεμένων συσκευών στο διαδίκτυο.

Το 2012 η κυβέρνηση Ομπάμα ανακοινώνει την πρωτοβουλία έρευνας και ανάπτυξης μεγάλων δεδομένων που αποτελείται απο 84 προγράμματα σε έξι τομείς. Το NSF δημοσιεύει «Βασικές τεχνικές και τεχνολογίες για την εξελισσόμενη επιστήμη και μηχανική μεγάλων δεδομένων». Τα IDC και EMC εκτιμούν οτι 2.8zettabytes δεδομένων θα δημιουργηθούν το 2012 αλλά μόνο το 3% όσων θα μπορούσαν να χρησιμοποιηθούν για μεγάλα δεδομένα είναι σημειωμένο και ακόμα λιγότερα εχουν αναλυθεί. Η αναφορά προβλέπει οτι μέχρι το 2020 ο ψηφιακός κόσμος θα κατέχει 40zettabytes, 57 φορές των συνολικό αριθμό των κόκκων άμμου απο όλες τις παραλίες του κόσμου.

Το «Harvard Business Review» αναφέρει το επάγγελμα του αναλυτή δεδομένων ως «την πιο σέξυ εργασία του 21ου αιώνα».

Το 2013 ο εκδημοκρατισμός των δεδομένων ξεκινά. Με έξυπνα τηλέφωνα, ταμπλέτες και ασύρματες συνδέσεις WiFi, όλοι πλέον παράγουν δεδομένα με ξέφρενους ρυθμούς. Ολοένα και περισσότερα άτομα έχουν πρόσβαση σε μεγάλους όγκους δημόσιων δεδομένων και αξιοποίησής τους με δημιουργική χρήση.

Τα γεγονότα των τελευταίων 20 ετών έχουν εκ βάθρων αλλάξει τον τρόπο με τον οποίο αντιμετωπίζονται τα δεδομένα. (Dean, 2014)

Κεφάλαιο 2ο

2.1 ΤΙ ΕΝΝΟΟΥΜΕ ΜΕ ΤΟΝ ΟΡΟ «ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ»

Ο όρος "Big Data" (μεγάλα δεδομένα) χρησιμοποιείται για να περιγράψει δεδομένα τα οποία χαρακτηρίζονται από εξαιρετικά μεγάλο όγκο, ο οποίος καθιστά ιδιαίτερα δύσκολη την εξόρυξη, αποθήκευση, διαχείριση και ανάλυση τους από τις παραδοσιακές εφαρμογές διαχείρισης βάσεων δεδομένων. Ωστόσο, ο ορισμός αυτός εμπεριέχει ένα υποκειμενικό όριο του ελάχιστου όγκου που πρέπει να έχουν τα δεδομένα, έτσι ώστε να μπορούν να θεωρηθούν "μεγάλα δεδομένα". Λαμβάνεται η υπόθεση ότι, καθώς η τεχνολογία εξελίσσεται μέσα στο χρόνο, ο όγκος των πακέτων δεδομένων, που χαρακτηρίζονται ως μεγάλα δεδομένα αυξάνεται επίσης. Επίσης, πρέπει να σημειωθεί ότι ο ορισμός μπορεί να διαφέρει ανά τομέα, ανάλογα με το είδος των λογισμικών, που είναι ευρέως διαθέσιμα και ποιά είναι τα συνήθη μεγέθη των πακέτων δεδομένων σε κάθε κλάδο. Με αυτές τις επισημάνσεις, τα μεγάλα δεδομένα σε πολλούς τομείς σήμερα κυμαίνονται από μερικές δεκάδες terabytes έως πολλαπλάσια petabytes (χιλιάδες terabytes).Ψηφιακά δεδομένα συναντώνται πλέον παντού: σε κάθε τομέα, σε κάθε οικονομία, σε κάθε οργανισμό και χρήστη της ψηφιακής τεχνολογίας. Τα μεγάλα δεδομένα έλκουν όλο και περισσότερο το ενδιαφέρον των ηγετών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται να ωφεληθούν από την αξιοποίησή τους. Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή. Τάσεις όπως ο Νόμος του Moore1 στην πληροφορική και το ισοδύναμό του στην ψηφιακή αποθήκευση και το cloud computing εξακολουθούν να μειώνουν κόστος και να εξαλείφουν τεχνολογικά εμπόδια. Με λιγότερο από 600 $,

[1] Ο Νόμος του Moore, ο οποίος περιγράφηκε για πρώτη φορά από τον συνιδρυτή της Intel Gordon Moore, αναφέρει ότι η πυκνότητα των τρανζίστορ στα τσιπ (ο αριθμός των τρανζίστορ ανά μονάδα επιφάνειας) διπλασιάζεται κάθε περίπου δύο χρόνια. Με άλλα λόγια, η ποσότητα υπολογιστικής ισχύος που μπορούν να αγοραστούν με το ίδιο χρηματικό ποσό διπλασιάζεται περίπου κάθε δύο χρόνια. Το cloud computing αναφέρεται στη δυνατότητα πρόσβασης σε υψηλή ποσότητα κλιμακούμενης (scalable) υπολογιστικής δύναμης μέσω του Διαδικτύου, συχνά σε τιμές χαμηλότερες από αυτές που θα απαιτούνταν για την εγκάταση στον υπολογιστή κάποιου , διότι οι πόροι διαμοιράζονται σε πολλούς χρήστες.

κάποιος μπορεί να αγοράσει μια μονάδα δίσκου με ικανότητα να αποθηκεύσει όλη τη μουσική του κόσμου. Επίσης, τα μέσα εξόρυξης γνώσης από τα δεδομένα σημειώνουν σημαντική βελτίωση, καθώς τα διαθέσιμα λογισμικά για την εφαρμογή τεχνικών αυξανόμενης πολυπλοκότητας συνδυάζονται με την αυξανόμενη υπολογιστική ισχύ. Επιπλέον, η δυνατότητα παραγωγής, επικοινωνίας, μερισμού και πρόσβασης δεδομένων έχει εκτοξευθεί από την αύξηση του αριθμού των ατόμων, συσκευών και αισθητήρων, που συνδέονται σήμερα σε ψηφιακά δίκτυα. Το 2010, περισσότερα από 4 δισεκατομμύρια άνθρωποι, ή το 60 τοις εκατό του παγκόσμιου πληθυσμού, χρησιμοποιούσαν κινητά τηλέφωνα, και περίπου 12 τοις εκατό από αυτούς τους ανθρώπους είχαν smartphones, των οποίων η διείσδυση αυξάνεται κατά περισσότερο από 20 τοις εκατό το χρόνο. Περισσότερα από 30 εκατ. δικτυωμένοι κόμβοι αισθητήρων βρίσκονται πλέον στους κλάδους μεταφορών, αυτοκινητοβιομηχανίας, επιχειρήσεων κοινής ωφέλειας, καθώς και σε τομείς του λιανικού εμπορίου. Ο αριθμός αυτών των αισθητήρων αυξάνεται σε ποσοστό άνω του 30 .(McKinsey, 2011)

Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή δεδομένων .Αυτός είναι ο λόγος που τα μεγάλης κλίμακας δεδομένων έχουν γίνει πρόσφατη περιοχή των στρατηγικών επενδύσεων για τους IT οργανισμούς .Αν και είναι σαφές ότι οι νέες τεχνολογίες και νέες μορφές προσωπικής επικοινωνίας οδήγησαν στην τάση των μεγάλης κλίμακας δεδομένων , θεωρούν ότι ο παγκόσμιος πληθυσμός του διαδικτύου αυξήθηκε κατά 6,5% από το 2010-2011 και τώρα αντιπροσωπεύει πάνω από δισεκατομμύρια ανθρώπους. Αυτό μπορεί να φαίνεται μεγάλο ,αλλά υποδηλώνει ότι η συντριπτική πλειοψηφία του παγκόσμιου ιστού έχει ακόμα να συνδεθεί .Ενώ μπορεί να είναι ότι ποτέ δεν θα φτάσουμε 100% του παγκόσμιου πληθυσμού σε απευθείας σύνδεση (λόγω των περιορισμένων διαθέσιμων πόρων , κόστος των αγαθών ,και την περιορισμένη υλική ευελιξία),όλο και περισσότερο είναι εκείνα που είναι σε απευθείας σύνδεση περισσότερο από ποτέ .Μόλις λίγα χρόνια πριν ήταν λογικό να σκεφτείς ότι πολλά είχαν μια επιφάνεια εργασίας (ίσως στην εργασία)και ίσως ένα laptop στην διάθεση τους. (Juniper networks, 2012)

- Το 2011, η ανθρωπότητα δημιούργησε πάνω από 1,2 τρισεκατομμύρια GB δεδομένων.
- Ο όγκος των δεδομένων αναμένεται να αυξηθεί 50 φορές μέχρι το 2020.

- Η Google λαμβάνει πάνω από 2.000.000 ερωτήματα αναζήτησης κάθε λεπτό.
- 72 ώρες βίντεο προστίθενται στο YouTube κάθε λεπτό.
- Υπάρχουν 217 νέοι χρήστες του Ιντερνέτ κάθε λεπτό.
- Οι χρήστες του Twitter στέλνουν πάνω από 100.000 tweets κάθε λεπτό (που είναι πάνω από 140 εκατομμύρια ανά ημέρα).
- Εταιρείες, και οργανισμοί να λάβουν 34.000 "likes" σε κοινωνικά δίκτυα κάθε λεπτό.
- Διεθνή δεδομένα Corporation (IDC) προβλέπει ότι η αγορά για την τεχνολογί α των μεγάλης κλίμακας δεδομένων και υπηρεσίες θα φτάσει τα $16,9
- δισεκατομμύρια μέχρι το 2015 με αύξηση 40% πάνω από τον ορίζοντα της πρόβλεψης. (Juniper networks, 2012)

2.2 Η ΣΠΟΥΔΑΙΌΤΗΤΑ ΤΩΝ ΜΕΓΆΛΩΝ ΔΕΔΟΜΈΝΩΝ

Η χρήση των μεγάλων δεδομένων προσφέρει τεράστιες ανεκμετάλλευτες δυνατότητες δημιουργικής αξίας. Οργανισμοί σε πολλούς κλάδους και πολλές επιχειρηματικές λειτουργίες μπορούν να αξιοποιήσουν μεγάλα δεδομένα με σκοπό τη βελτίωση της κατανομής και του συντονισμού των πόρων τους, τον περιορισμό της σπατάλης, την αύξηση της διαφάνειας και της λογοδοσίας, και την ανάδειξη νέων ιδεών και αντιλήψεων.

Τα μεγάλα δεδομένα δημιουργούν αξία με διάφορους τρόπους. Οι σημαντικότεροι είναι οι κάτωθι:

Δημιουργία Διαφάνειας (transparency)

Η εύκολη και έγκαιρη πρόσβαση σε μεγάλα δεδομένα από τους ενδιαφερόμενους φορείς παρέχει ευκαιρίες δημιουργίας τεράστιας αξίας. Συχνά, τέτοιες ευκαιρίες προκύπτουν σε περιπτώσεις όπου παρατηρείται έλλειψη συμφωνίας κινήτρων για δημιουργία διαφάνειας δεδομένων. Για παράδειγμα, στον δημόσιο τομέα, υπάρχουν περιπτώσεις όπου το προσωπικό διαφόρων υπηρεσιών σπαταλά σημαντικό ποσοστό του χρόνου τους για να εντοπίσουν πληροφορίες σε άλλες κυβερνητικές υπηρεσίες, χρησιμοποιώντας μη-ψηφιακά μέσα (π.χ. σε έντυπους καταλόγους ή τηλεφωνώντας), και στη συνέχεια για να πάρουν τις πληροφορίες αυτές έπρεπε να επισκεφθούν την πηγή της πληροφορίας για να λάβουν τα στοιχεία με φυσικά μέσα, (π.χ οπτικοί δίσκοι). Τέτοιου είδους σπατάλη έχει μειωθεί σημαντικά σε οργανισμούς, που αξιοποιούν τα μεγάλα δεδομένα για να ψηφιοποιήσουν την πληροφορία αυτή, χρησιμοποιώντας τα διαθέσιμα δίκτυα, και αναπτύσσοντας εργαλεία ευκολότερης εύρεσης της αναζητούμενης πληροφορίας.

Ωστόσο, ακόμη και σε τομείς, που έχουν υιοθετηθεί οι νέες τεχνολογίες και τα μεγάλα δεδομένα, υπάρχουν σημαντικά κίνητρα για υψηλότερη απόδοση, υπάρχουν περιθώρια αύξησης διαφάνειας και ανταλλαγής μεγάλων δεδομένων. Στον τομέα της μεταποίησης, πολλές εταιρείες χρησιμοποιούν τα μεγάλα δεδομένα για τη βελτίωση στην απόδοση της Έρευνας και Τεχνολογίας (π.χ. πολύπλοκες προσομοιώσεις) και στη διαχείριση της αλυσίδας εφοδιασμού τους. (McKinsey, 2011)

Εντοπισμός Αναγκών, Μεταβλητότητας, και Άυξηση Απόδοσης

Όλο και περισσότερες εταιρείες ψηφιοποιούν και αποθηκεύουν μια αυξανόμενη ποσότητα εξαιρετικά λεπτομερών δεδομένων σχετικά με τις συναλλαγές. Όλο και περισσότεροι αισθητήρες ενσωματώνονται σε φυσικές συσκευές - από τον εξοπλισμό της γραμμής παραγωγής έως σε αυτοκίνητα και σε κινητά τηλέφωνα- οι οποίοι μετρούν διαδικασίες, χρήση προϊόντων, και ανθρώπινες συμπεριφορές. Επίσης, ατομικά οι καταναλωτές δημιουργούν και μοιράζονται μια τεράστια ποσότητα δεδομένων μέσω του blogging, των ενημερώσεων κατάστασης, και την ανάρτηση φωτογραφίων και βίντεο. Μεγάλο μέρος των δεδομένων αυτών μπορεί τώρα να συγκεντρώνεται σε πραγματικό ή σχεδόν πραγματικό χρόνο.

Η δυνατότητα πρόσβασης σε όλα τα δεδομένα αυτά και, σε ορισμένες περιπτώσεις, η δυνατότητα διαχείρισης των συνθηκών δημιουργίας τους, παρέχουν έναν πολύ διαφορετικό τρόπο λήψης αποφάσεων, τον οποίο εισάγει πιο πολύ η επιστήμη στη Διοίκηση. Πιό συγκεκριμένα, οι μάνατζερς μπορούν να χρησιμοποιήσουν τώρα επιστημονικές διαδικασίες ελεγχόμενης έρευνας, που περιλαμβάνουν τον ορισμό συγκεκριμένων υποθέσεων, τον σχεδιασμό και την εκπόνηση ερευνών για να επαληθεύσουν τις υποθέσεις αυτές, και στη συνέχεια να αναλύσουν διεξοδικά τα ποσοτικά ευρήματα πριν από τη λήψη απόφασης. Ένας οργανισμός, που είναι προσανατολισμένος στα δεδομένα λαμβάνει αποφάσεις με βάση τα εμπειρικά αποτελέσματα, και τα οφέλη μιας τέτοιας προσέγγισης έχουν αποδειχθεί και από την ακαδημαϊκή έρευνα.

Οι ηγέτες σε πολλούς τομείς έχουν ήδη αρχίσει να χρησιμοποιούν ελεγχόμενες έρευνες για τη λήψη καλύτερων αποφάσεων. Για παράδειγμα, στον τομέα της υγείας εκπονούνται μελέτες συγκριτικής αξιολόγησης αποτελεσματικότητας σε ολόκληρο τον πληθυσμό, καθώς εντοπίζονται επαρκή κλινικά δεδομένα για τον εντοπισμό και την κατανόηση των πηγών της μεταβλητότητας σε θεραπείες και αποτελέσματα και έτσι βοηθούνται οι υπεύθυνοι για τη λήψη αποφάσεων στη χάραξη κατευθυντήριων γραμμών, που εξασφαλίζουν ότι οι αποφάσεις για τη θεραπεία βασίζονται στην ορθότερη επιστήμη. Οι πωλητές, κυρίως εκείνοι, που δραστηροποιούνται διαδικτυακά, αλλά διαρκώς και εκείνοι με τα φυσικά καταστήματα,

προσαρμόζουν τιμές και προσφορές σε μια προσπάθεια εντοπισμού του βέλτιστου συνδυασμού κυκλοφορίας και πωλήσεων.

Ωστόσο, δεν είναι πάντα δυνατόν (για λόγους ηθικής ή εφικτότητας) η κατασκευή μιας ελεγχόμενης έρευνας και ο «χειρισμός» μια ανεξάρτητης μεταβλητής. Μια εναλλακτική είναι η εύρεση «φυσικών πειραμάτων», που εντοπίζουν την υπάρχουσα μεταβλητότητα στις μετρήσεις απόδοσης. Η κατανόηση των αιτιών αυτής της μεταβλητότητας μπορεί στη συνέχεια να συμβουλέψει τους υπευθύνους διαχείρισης να λάβουν αποφάσεις και να βελτιώσουν την απόδοση. Στο δημόσιο τομέα, εντοπίζονται υπηρεσίες με τεράστιες αποκλίσεις στην παραγωγικότητα και την ακρίβεια του έργου, οι οποίες εκτελούν σχεδόν πανομοιότυπα καθήκοντα. Η γνωστοποίηση και μόνο αυτής της πληροφορίας μπορεί να έχει ως αποτέλεσμα σημαντική αύξηση απόδοσης στις υστερούσες υπηρεσίες και χωρίς χρηματικό αντίκρυσμα ως κίνητρο. (McKinsey, 2011)

Κατάτμηση του πληθυσμού για την προσαρμογή δράσεων

Οι πολιτικές στοχευμένων υπηρεσιών ή του μάρκετινγκ για να ανταποκρίνονται στις ανάγκες των ατόμων είναι ήδη οικείες σε εταιρείες προσανατολισμένες προς την ιδιωτική κατανάλωση. Η ιδέα της κατάτμησης της αγοράς και της ανάλυσης των πελατών τους μέσω συνδυασμών χαρακτηριστικών όπως δημογραφικά στοιχεία, μετρήσεις αγορών πελατών, και αγοραστικές συμπεριφορές είναι ευρέως καθιερωμένες. Επιχειρήσεις όπως οι ασφαλιστικές εταιρείες, οι οποίες βασίζονται σε αποφάσεις αβεβαιότητας έχουν χρησιμοποιήσει επί μακρόν μεγάλα δεδομένα για την τμηματοποίηση. Ωστόσο, καθώς η τεχνολογία εξελίσσεται, πολλές εταιρείες αποκτούν τη δυνατότητα να τμηματοποιούν και να αναλύουν σε πραγματικό χρόνο. Ακόμη και στο δημόσιο τομέα, που η τάση είναι να αντιμετωπίζονται όλες οι δομές με τον ίδιο τρόπο, η χρήση μεγάλων δεδομένων για τμηματοποίηση αρχίζει να εφαρμόζεται. Για παράδειγμα, οι φορολογικές υπηρεσίες όπου οι φορολογούμενοι τμηματοποιούνται από μια σειρά παραγόντων όπως το εισόδημα, το ποσοστό φερεγγυότητάς τους και το πιστωτικό ιστορικό τους για την επιλογή μέσων κατάλληλων για περαιτέρω ελέγχους. (McKinsey, 2011)

Αντικατάσταση / υποστήριξη της λήψης αποφάσεων με αυτοματοποιημένους αλγόριθμους

Εξελιγμένα analytics μπορεί να βελτιώσουν σημαντικά τη λήψη αποφάσεων, την ελαχιστοποίηση της αβεβαιότητας, και την ανάδειξη πολύτιμων πληροφορίων. Τα μεγάλα δεδομένα παρέχουν την πρώτη ύλη, που απαιτείται είτε για την ανάπτυξη αλγορίθμων, είτε για τη λειτουργία τους. Για παράδειγμα, φορολογικές υπηρεσίες, που εφαρμόζουν και χρησιμοποιούν αυτοματοποιημένες μηχανές αβεβαιότητας που χρησιμοποιούν μεγάλα δεδομένα για τον εντοπισμό υποψηφίων, που χρήζουν περαιτέρω διερεύνησης. Οι αλγόριθμοι μεγάλων δεδομένων στον τομέα της λιανικής μπορούν να αριστοποιήσουν τις

διαδικασίες λήψης αποφάσεων, επιτρέποντας την αυτόματη ρύθμιση καταλόγων και τιμολογώντας σε πραγματικό χρόνο και σε κατάστηματα και σε online πωλήσεις. Οι κατασκευαστικές εταιρείες μπορούν να προσαρμόσουν τις γραμμές παραγωγής τους αυτόματα, για βελτιστοποίηση αποδοτικότητας, μείωση σπατάλης, και αποφυγή επικίνδυνων συνθηκών. Σε ορισμένες περιπτώσεις, εταιρείες δεν αυτοματοποιούν απαραίτητα τις αποφάσεις, αλλά τις διευκολύνουν μέσω της ανάλυσης των μεγάλων δεδομένων, που είναι πολύ περισσότερα από τα δεδομένα, που είναι διαχειρίσιμα από ένα άτομο χρησιμοποιώντας ένα υπολογιστικό φύλλο. Ορισμένοι οργανισμοί λαμβάνουν ήδη πιο αποτελεσματικές αποφάσεις αναλύοντας ολόκληρα σύνολα δεδομένων από πελάτες και εργαζόμενους, ή ακόμα και από αισθητήρες ενσωματωμένους σε προϊόντα. (McKinsey, 2011)

Καινοτόμα νέα επιχειρηματικά μοντέλα, προϊόντα και υπηρεσίες

Τα μεγάλα δεδομένα επιτρέπουν στις επιχειρήσεις όλων των ειδών την ανάπτυξη νέων προϊόντων και υπηρεσιών, την ενίσχυση των υφιστάμενων, και την εισαγωγή εντελώς νέων επιχειρηματικών μοντέλων. Στον τομέα της Υγείας, η ανάλυση των κλινικών δεδομένων και δεδομένων τη συμπεριφοράς των ασθενών έχει οδηγήσει σε προγράμματα προληπτικής φροντίδας, στοχευόμενα στις κατάλληλες ομάδες ατόμων. Η εταιρεία Ingenix στον τομέα της υγειονομικής περίθαλψης και η Nielsen στο λιανικό εμπόριο ειδικεύονται στη συγκέντρωση και ανάλυση των συνόλων δεδομένων για διάφορα ιδρύματα. Επίσης, στο λιανικό εμπόριο, οι υπηρεσίες σύγκρισης τιμών σε πραγματικό χρόνο δίνουν στους καταναλωτές πλήρη εικόνα των τιμών σε βαθμό, που ποτέ πριν δεν απολάμβαναν και δημιουργούν σημαντικό πλεόνασμα για αυτούς. Άλλες εταιρείες χρησιμοποιούν δεδομένα που λαμβάνονται από αισθητήρες ενσωματωμένους σε προϊόντα για τη δημιουργία καινοτόμων μετά την πώληση προσφορών υπηρεσιών, όπως η προληπτική συντήρηση και για τη δημιουργία βάσης για την ανάπτυξη της επόμενης γενιάς προϊόντων. (McKinsey, 2011)

2.3 ΠΕΡΙΠΤΩΣΕΙΣ ΧΡΗΣΗΣ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ ΔΕΔΟΜΕΝΩΝ

Υπάρχουν πολλά παραδείγματα περιπτώσεων χρήσης των μεγάλων κλίμακας δεδομένων σε κάθε βιομηχανία που μπορεί να φανταστεί κανείς. Ορισμένες επιχειρήσεις έχουν γίνει πιο δεκτικές στις τεχνολογίες και έχουν ενσωματώσει πιο γρήγορα την ανάλυση δεδομένων στην καθημερινότητα της επιχείρησης σε σχέση με άλλες. Αυτό είναι προφανές ότι οι επιχείρησης που αγκαλιάζουν την τεχνολογία όχι μόνο θα δουν σημαντικά πρωτοποριακά πλεονεκτήματα ,αλλά θα είναι σημαντικά πιο ευέλικτες και πιο προσαρμοστικές στις προσφορές τους.

Παραδείγματα χρήσης των μεγάλων κλίμακας δεδομένων περιλαμβάνουν:

1. Οι χρηματοπιστωτικές υπηρεσίες υιοθετούν υποδομές ανάλυσης μεγάλων δεδομένων για να βελτιώσουν τις αναλύσεις των πελατών τους για το μετοχικό κεφάλαιο, ασφάλιση, υποθήκη, ή πίστωση. (Juniper networks, 2012)

2. Αεροπορικές εταιρείες και εταιρείες οδικών μεταφορών χρησιμοποιούν μεγάλης κλίμακας δεδομένων για να παρακολουθήσουν την κατανάλωση καύσιμων και τα πρότυπα κυκλοφορίας στους στόλους τους σε πραγματικό χρόνο για να βελτιώσουν την αποτελεσματικότητα και την εξοικονόμηση κόστους. (Juniper networks, 2012)

3. Οι υγειονομικής περίθαλψης υπηρεσίες διαχειρίζονται και κάνουν κοινή χρήση ηλεκτρονικών μητρώων ασθενών από πολλαπλές πήγες —εικόνες, θεραπείες, και δημογραφικά στοιχεία. Επιπλέον, οι φαρμακευτικές εταιρείες και οι ρυθμιστικοί οργανισμοί δημιουργούν λύσεις μεγάλης κλίμακας δεδομένων για την παρακολούθηση της αποτελεσματικότητας των φαρμάκων και για να παρέχουν πιο αποτελεσματική και πιο σύντομη ανάπτυξη φαρμάκων. (Juniper networks, 2012)

4. Εταιρίες μέσων ενημέρωσης και ψυχαγωγίας αξιοποιούν τις υποδομές της μεγάλης κλίμακας δεδομένων για να βοηθήσουν με την λήψη αποφάσεων γύρω από τον πελάτη και για να παρέχει πιο εστιασμένο μάρκετινγκ. (Juniper networks, 2012)

Υπάρχουν περιπτώσεις χρήσης και συγκεκριμένα παραδείγματα των μεγάλης κλίμακας δεδομένων για κάθε βιομηχανία και εταιρεία . Ως εκ τούτου, έστω και αν αυτήν την περίοδο η επιχείρηση σας δεν χρησιμοποιεί λύσεις μεγάλων δεδομένων, είναι πιθανόν οι ανταγωνιστές σας να χρησιμοποιούν. Το πραγματικό ερώτημα είναι πως μπορείς να βελτιστοποιήσεις καλυτέρα το περιβάλλον σου ώστε να δημιουργήσεις μια πιο γρήγορη αποτελεσματική λύση που σου δίνει ανταγωνιστικό πλεονέκτημα.

Συμφώνα με την ερευνά από McKinsey Global Institute(MGI),αναλύοντας τα μεγάλα σύνολα δεδομένων θα γίνει –και έχει ήδη γίνει για ένα μεγάλο αριθμό επιχειρήσεων –ένα εργαλείο σχεδιασμού .Με την επιφύλαξη ότι με τις σωστές πολιτικές και προϋποθέσεις που πρέπει να εφαρμοστούν ,τα μεγάλης κλίμακας δεδομένων θα γίνουν ένα κρίσιμο εργαλείο για την ανάπτυξη σχεδίων για:

- Ανταγωνιστικό σχεδιασμό και έρευνα
- Μελλοντική παραγωγικότητα και προϊόν ανάπτυξης
- Καινοτομία προϊόντων και υπηρεσιών
- Ικανοποίηση του πελάτη (όπως ορίζεται στη μελέτη ,"Πλεόνασμα καταναλωτών") (Juniper networks, 2012)

2.4. ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΣΗΜΑΝΤΙΚΟΤΕΡΩΝ ΟΡΙΣΜΩΝ ΓΙΑ ΤΑ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

Από το 2011 το ενδιαφέρον για το χώρο των μεγάλων δεδομένων έχει αυξηθεί με εκθετικό βαθμό. Σε αντίθεση με την συντριπτική πλειοψηφία των ερευνών σχετικά με την επιστήμη των υπολογιστών, τα Μεγάλα δεδομένα έλαβαν μεγάλη δημοσιότητα και ενδιαφέρον από τα μέσα ενημέρωσης. Τίτλοι όπως "Μεγάλα δεδομένα: το μεγαλύτερο αγαθό ή καταπάτηση της ιδιωτικότητας (P.Chatterjee, 2013) " και "Τα Μεγάλα δεδομένα ανοίγουν πόρτες αλλά ίσως πάρα πολλές (Big Data Is Opening Doors, but Maybe Too Many, 2013)" λένε πολλά ως προς την αντίληψη που επικρατεί για τα μεγάλα δεδομένα. Από την αρχή γίνεται σαφές ότι τα Μεγάλα δεδομένα σχετίζονται με σημαντικά τεχνικά αλλά και κοινωνικό-τεχνικά θέματα αλλά ο ακριβής ορισμός τους δεν είναι αρκετά σαφής. Πρόσφατη βιβλιογραφία που κάνει χρήση του όρου συναντάται σε πολλά και διαφορετικά πεδία με αποτέλεσμα την ύπαρξη πολλών, διφορούμενων και συχνά αντιφατικών ορισμών σχετικά με τον όρο Μεγάλα δεδομένα.

Τα Μεγάλα δεδομένα σχετίζονται κυρίως με δύο ιδέες: την αποθήκευση δεδομένων και την ανάλυση δεδομένων. Σε αντίθεση με αυτό το ξαφνικό ενδιαφέρον για τα μεγάλα δεδομένα, οι έννοιες αυτές δεν είναι καινούργιες στον επιστημονικό κόσμο. Αυτό στο στοιχείο ωστόσο, αναδεικνύει το ερώτημα του πώς τα Μεγάλα δεδομένα θεωρούνται σημαντικά διαφορετικά από τις τυπικές τεχνικές επεξεργασίας δεδομένων. Δεν χρειάζεται ιδιαίτερη διορατικότητα για να καταλάβουμε ότι για να βρούμε την απάντηση σε αυτό το ερώτημα πρέπει απλώς να εξεταστεί περαιτέρω ο όρος μεγάλο δεδομένα. Ο όρος "Μεγάλα" υποδηλώνει σημαντικότητα, πολυπλοκότητα και πρόκληση. Δυστυχώς όμως ο όρος "Μεγάλα" περιέχει ποσοτικό χαρακτηριστικό και εδώ έγκειται η δυσκολία για την εξαγωγή ενός ορισμού.

Ένας εκ των πιο διαδεδομένων ορισμών περιλαμβάνεται σε έκθεση του Meta (σήμερα Gartner) το 2001 (3d data management:Controlling data volume,velocity and variety, 2001). Η έκθεση της Gartner δεν κάνει καμία αναφορά στη φράση «μεγάλα δεδομένα», ωστόσο, η έκθεση αυτή θεωρείτε βασικός ορισμός των Μεγάλα δεδομένα. Η Gartner πρότεινε έναν ορισμό που περιλάμβανε τα "τρία Vs (Volume, Velocity, Variety)": τον όγκο, την ταχύτητα και την ποικιλία. Πρόκειται για έναν ορισμό που εστιάζει στο μέγεθος. Η έκθεση επισημαίνει το αυξανόμενο μέγεθος των δεδομένων, το αυξανόμενο ποσοστό παραγωγής τους και το αυξανόμενο εύρος των μορφών που εφαρμόζονται. Όπως είναι σύνηθες στη βιβλιογραφία των μεγάλων δεδομένων, τα ευρήματα που παρουσιάζονται στον ορισμό της Gartner είναι εντελώς αποσπασματικά και δεν παρέχεται καμία ποσοτικοποίηση των μεγάλων δεδομένων. Ο ορισμός αυτός έχει επαναληφθεί από τη NIST (Nist Big Data program, 2013) και τη Gartner το 2012 (M.A Beyer and D.Laney, 2012) και διευρυνθεί από την IBM (IBM, 2013) για να συμπεριλάβει και ένα τέταρτο V: την πιστότητα (Veracity).

Η Oracle αποφεύγει την χρήση των Vs για να καταλήξει σε έναν ορισμό. Αντ 'αυτού η Oracle (J.P Dijcks ORACLE, 2013) υποστηρίζει ότι τα μεγάλα στοιχεία είναι η δημιουργία αξίας από παραδοσιακές σχεσιακές βάσεις δεδομένων με στόχο τη λήψη επιχειρηματικών αποφάσεων, η οποία είναι εμπλουτισμένη με νέες πηγές μη-δομημένων δεδομένων. Οι νέες αυτές πηγές περιλαμβάνουν blogs, social media, δίκτυα αισθητήρων, δεδομένα εικόνας και άλλες μορφές δεδομένων, τα οποία ποικίλλουν σε μέγεθος, δομή, μορφή και άλλους παράγοντες. Η Oracle υποστηρίζει ότι τα μεγάλα δεδομένα είναι το αποτέλεσμα από την ένταξη πρόσθετων πηγών δεδομένων για να αυξήσουν τις ήδη υπάρχουσες λειτουργίες. Αξίζει να σημειωθεί ότι ο ορισμός της Oracle εστιάζει στην υποδομή. Σε αντίθεση με ορισμούς που εκφράστηκαν από άλλους, η Oracle δίνει έμφαση σε μια σειρά από τεχνολογίες όπως: NoSQL, Hadoop, HDFS, R και σχεσιακές βάσεις δεδομένων. Έτσι, παρείχαν και έναν ορισμό και μια λύση για τα μεγάλα δεδομένα. Παρόλο που ο ορισμός αυτός είναι σχετικά πιο εύκολο να υιοθετηθεί σε σχέση με άλλους, υστερεί ωστόσο στην ποσοτικοποίηση. Σύμφωνα με τον ορισμό της Oracle δεν είναι σαφές ως προς το πότε ακριβώς ο όρος μεγάλα δεδομένα εντοπίζεται στην πράξη και παρέχει περισσότερο μία έννοια ότι «θα τα καταλάβετε όταν τα δείτε".

Η Intel είναι μία από τις λίγες επιχειρήσεις που παρέχουν ποσοτικά στοιχεία στη βιβλιογραφία τους. Η Intel συσχετίζει τα μεγάλα δεδομένα με οργανισμούς που “δημιουργούν κατά μέσο όρο 300 terabytes (TB) δεδομένων εβδομαδιαίως". Αντί να δώσει έναν ορισμό όπως έκαναν οι προαναφερθέντες οργανισμοί, περιγράφει τα μεγάλα δεδομένα ποσοτικοποιώντας τις εμπειρίες των επιχειρηματικών εταίρων της. Επισημαίνει ότι οι οργανισμοί οι οποίοι μελετήθηκαν ασχολούνται εκτενώς με μη-δομημένα δεδομένα και δίνουν έμφαση στη διεξαγωγή αναλύσεων των δεδομένων τους τα οποία παράγονται με ρυθμό 500 terabytes ανά εβδομάδα. Τέλος, ισχυρίζεται ότι ο πιο σύνηθες τύπος δεδομένων που συναντάται είναι οι επιχειρηματικές συναλλαγές που είναι αποθηκευμένες σε σχεσιακές βάσεις δεδομένων (σύμφωνα με τον ορισμό της Oracle), και ακολουθούν τα έγγραφα, τα e-mail, τα blogs και τα social media.

Η Microsoft παρέχει ένα ιδιαίτερα περιεκτικό ορισμό: "Μεγάλα δεδομένα” είναι ο όρος που χρησιμοποιείται όλο και περισσότερο για να περιγράψει τη διαδικασία εφαρμογής σημαντικής υπολογιστικής ισχύς - την τελευταία λέξη της μηχανικής μάθησης και της τεχνητής νοημοσύνης - σε μαζικά και εξαιρετικά πολύπλοκα σύνολα πληροφοριών.Ο ορισμός αυτός καθιστά σαφές ότι τα μεγάλα δεδομένα απαιτούν σημαντική υπολογιστική ισχύ. Η σημασία της υπολογιστικής ισχύς αναφέρθηκε και σε προηγούμενους ορισμούς, αλλά δεν ορίστηκε με ακρίβεια. Επιπλέον, ο ορισμός αυτός εισάγει δύο τεχνολογίες: την μηχανική μάθηση και την τεχνητή νοημοσύνη που είχαν αγνοηθεί από προηγούμενους ορισμούς. Αυτό, ως εκ τούτου, εισάγει την ιδέα ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες που είναι ζωτικής σημασίας συστατικά του τελικού ορισμού.

Η Google Trends αναφέρει τους ακόλουθους όρους σε σχέση με τα μεγάλα δεδομένα2: ανάλυση δεδομένων, Hadoop, NoSQL, Google, IBM, και Oracle. Από αυτούς τους όρους μια σειρά από τάσεις είναι εμφανείς. Πρώτον, ότι τα μεγάλα δεδομένα είναι άρρηκτα συνδεδεμένα με την ανάλυση δεδομένων και την εξαγωγή γνώσης του από τα δεδομένα. Δεύτερον, είναι σαφές ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες όπως φαίνεται και από τον το ορισμό της Microsoft, δηλαδή τις NoSQL και Apache Hadoop. Τέλος, είναι προφανές ότι υπάρχει ένας αριθμός οργανισμών, κυρίως βιομηχανικών οργανισμών που σχετίζονται με μεγάλα δεδομένα.

Όπως επισημαίνεται από το Google Trends, υπάρχουν μια σειρά από τεχνολογίες που συχνά αναφέρονται ότι εμπλέκονται με τα μεγάλα δεδομένα. Αποθήκες δεδομένων NoSQL όπως Amazon, Dynamo, Cassandra, Couch DB, Mongo DB κ.ά. παίζουν κρίσιμο ρόλο στην αποθήκευση μεγάλου όγκου μη δομημένων και ιδιαίτερα μεταβαλλόμενων δεδομένων. Για τη χρήση των χώρων αποθήκευσης δεδομένων NoSQL υπάρχει μια σειρά εργαλείων ανάλυσης και μέθοδο, συμπεριλαμβανομένων των MapReduce, NLP, στατιστικού προγραμματισμού, της μηχανικής μάθησης και την οπτικοποίηση πληροφοριών. Η εφαρμογή μίας από αυτές τις τεχνολογίες από μόνη της δεν είναι επαρκής για να αξιολογήσει τη χρήση του όρου μεγάλα δεδομένα. Αντίθετα, άλλες τάσεις δείχνουν ότι είναι ο συνδυασμός μιας σειράς τεχνολογιών και η χρήση σημαντικών συνόλων δεδομένων που εξηγούν τον όρο. Οι τάσεις αυτές δείχνουν τα μεγάλα δεδομένα σαν μια τεχνική κίνηση η οποία ενσωματώνει ιδέες, νέες και παλιές και σε αντίθεση με άλλους ορισμούς παρέχει λίγες αναφορές ως προς τις κοινωνικές και επιχειρηματικές επιπτώσεις.

Καθώς οι προαναφερθέντες ορισμοί βασίζονται σε ένα συνδυασμό μεγέθους, πολυπλοκότητας και τεχνολογίας, ένα λιγότερο κοινός ορισμός βασίζεται μόνο στην πολυπλοκότητα. Η μέθοδος για ένα Ολοκληρωμένο σε Γνώση Περιβάλλον- (The Method for an Integrated Knowledge Environment (MIKE2.0) αποδίδεται συχνά στην κοινότητα ανοικτού κώδικα, εισάγοντας μια αντιφατική ιδέα: "Τα Μεγάλα δεδομένα μπορεί να είναι πολύ μικρά και δεν είναι όλα τα σύνολα δεδομένων μεγάλα" (Roberd Hillard, 2012). Αυτό είναι ένα επιχείρημα υπέρ της πολυπλοκότητας και υπέρ της άποψης ότι το μέγεθος δεν είναι το κυρίαρχο στοιχείο. Το πρόγραμμα MIKE υποστηρίζει ότι είναι ένας υψηλός βαθμός μεταθέσεων και αλληλεπιδράσεων μέσα σε ένα σύνολο δεδομένων που ορίζεται ως μεγάλα δεδομένα.

Η ιδέα εκφράστηκε στον ορισμό MIKE: Αυτά τα μεγάλα δεδομένα δεν είναι εύκολα διαχειρίσιμα με συμβατικά εργαλεία. Αυτή η ιδέα υποστηρίζεται από τον ορισμό του NIST ο οποίος αναφέρει ότι τα μεγάλα δεδομένα είναι δεδομένα τα οποία: "υπερβαίνουν την χωρητικότητα ή την ικανότητα των σημερινών ή συμβατικών μεθόδων και συστημάτων"

[2] Google. Google Trends for Big Data, 2013.

(Nist Big Data program, 2013)3. Δεδομένης της συνεχώς αναπτυσσόμενης φύσης της επιστήμης των υπολογιστών ο ορισμός αυτός δεν είναι τόσο πολύτιμος όσο μπορεί να φαινόταν αρχικά. Ο ισχυρισμός ότι τα μεγάλα δεδομένα είναι δεδομένα που αμφισβητούν και προκαλούν τις υφιστάμενες πρακτικές δεν είναι κάτι νέο. Ο ορισμός αυτός υποδηλώνει ότι τα δεδομένα είναι "μεγάλα" σε σχέση με το ισχύον πρότυπο υπολογισμού. Η εφαρμογή πρόσθετων υπολογισμών ή ακόμη η ανάπτυξη των υπαρχόντων υπόσχεται να συρρικνώσει τα μεγάλα δεδομένα. Ο ορισμός αυτός μπορεί να χρησιμεύσει μόνο ως ένα σύνολο από συνεχώς ανανεωνόμενων στόχων και υποστηρίζει ότι τα μεγάλα δεδομένα υπήρχαν ανέκαθεν, και πάντα θα υπάρχουν.

Παρά το εύρος και τις διαφορές που υπάρχουν σε καθένα από τους προαναφερθέντες ορισμούς υπάρχουν μερικά σημεία ομοιότητας. Αξίζει να σημειωθεί ότι όλοι οι ορισμοί κάνουν τουλάχιστον ένα από τα παρακάτω ισχυρισμούς:

Μέγεθος: ο όγκος των συνόλων δεδομένων είναι ένας κρίσιμος παράγοντας.

Πολυπλοκότητα: η δομή, η συμπεριφορά και οι μεταθέσεις των συνόλων δεδομένων είναι ένας κρίσιμος παράγοντας.

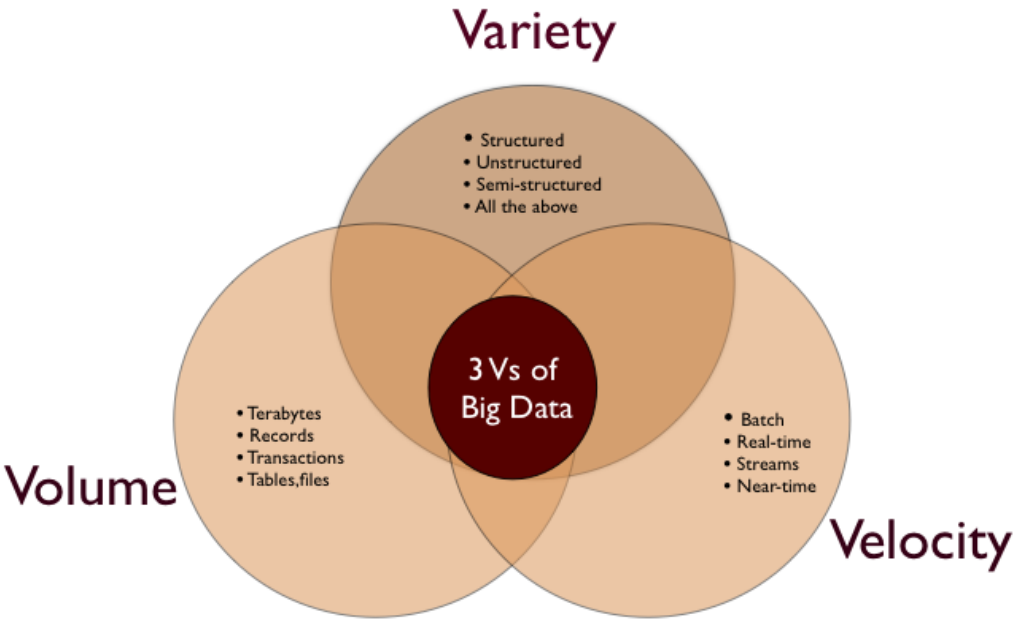
Τεχνολογίες: τα εργαλεία και οι τεχνικές που χρησιμοποιούνται για την επεξεργασία ενός μεγάλου και πολύπλοκου συνόλου δεδομένων είναι ένας κρίσιμος παράγοντας.

Οι ορισμοί που αναφέραμε παραπάνω περιλαμβάνουν τουλάχιστον έναν από αυτούς τους παράγοντες και πολλοί από αυτούς περιλαμβάνουν δύο. Μια προέκταση αυτών των παραγόντων θα ήταν επομένως να υποθέσουμε τα εξής: Μεγάλα δεδομένα είναι ένας όρος που περιγράφει την αποθήκευση και ανάλυση μεγάλων ή πολύπλοκων συνόλων δεδομένων χρησιμοποιώντας μια σειρά από τεχνικές που περιλαμβάνουν, αλλά δεν περιορίζονται στις ακόλουθες: NoSQL, MapReduce και μηχανές μάθησης.

2.5 ΤΟ ΠΡΟΤΥΠΟ ΤΩΝ 3V

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, ένας από τους πιο γνωστούς ορισμούς για τα Μεγάλα Δεδομένα διατυπώθηκε πολύ εύστοχα από την Gartner με βάση το αποκαλούμενο πρότυπο 3V. Σύμφωνα με αυτή την προσέγγιση, τρία είναι τα κύρια χαρακτηριστικά των μεγάλων δεδομένων: ο όγκος (Volume), η ταχύτητα (Velocity) και η ποικιλομορφία (Variety). Ο όγκος αναφέρεται στο μέγεθος των διαθέσιμων δεδομένων, η ποικιλία αναφέρεται στο μεγάλο εύρος διαφορετικών τύπων δεδομένων που πρέπει να διαχειριστούμε και η ταχύτητα αναφέρεται στον ρυθμό που τα δεδομένα παράγονται και επεξεργάζονται.

[3] NIST Big Data Working Group (NBD-WG). http://bigdatawg.nist.gov/home.php.



Σχήμα 1: Το πρότυπο των 3V των μεγάλων δεδομένων (https://kavyamuthanna.files.wordpress.com/2013/01/picture-11.png)

Όγκος (Volume)

Τα τελευταία χρόνια παρατηρούμε μια εντυπωσιακή αύξηση του όγκου των δεδομένων που καλούμαστε να αποθηκεύσουμε. Διαμορφώνεται ένα νέο περιβάλλον όπου δεν κυριαρχούν πια τα δεδομένα κειμένου. Αυτή η νέα πραγματικότητα, χαρακτηρίζεται από τεράστιες ποσότητες δεδομένων σε μορφή video, ήχου και εικόνων όχι μόνο επιστημονικής προελεύσεως αλλά έχοντας αξιοσημείωτη πηγή τα μέσα κοινωνικής δικτύωσης. Είναι απαραίτητο πια στοιχείο για έναν οργανισμό επιχείρηση να έχει στη διάθεσή του μεγάλο αποθηκευτικό χώρο. Εκ των πραγμάτων, η τεράστια αύξηση των δεδομένων, επιτάσσει την επανεξέταση εφαρμογών και αρχιτεκτονικών καθώς οι τυπικές μέθοδοι αποδεικνύονται ανεπαρκείς. Ξεπερνώντας σιγά σιγά τα προβλήματα εύρεσης επαρκούς χώρου αποθήκευσης, νέα ζητήματα αναδύονται όπως η ανάγκη συσχέτισης των μεγάλων δεδομένων και η δυνατότητα αλίευσης πολύτιμης αξίας.

Ταχύτητα (Velocity)

Καθώς τα δεδομένα ρέουν με καταιγιστικό ρυθμό αναδύεται η ανάγκη άμεσης αντίδρασής μας σε αυτά. Η ταχεία αντίδραση μας ώστε να αντιμετωπίσουμε την ταχύτητα των δεδομένων αποτελεί ιδιαίτερη πρόκληση για τους περισσότερους οργανισμούς. Η ταχύτητα αναφέρεται στον πολύ γρήγορο ρυθμό εμφάνισης νέων δεδομένων αλλά και ανανέωσης των υπαρχόντων. Επίσης, σχετίζεται με τον αναγκαίο χρόνο επεξεργασίας των εισερχομένων στο σύστημα

δεδομένων μέσω προηγμένων εφαρμογών, με την ανάλυση των δεδομένων αυτών, τον εντοπισμό των σχέσεων μεταξύ δεδομένων και την εξαγωγή πληροφοριών από τα δεδομένα μέσω συσχετίσεων και συμπερασμάτων. Αναφορικά με το ρυθμό εμφάνισης νέων δεδομένων, ο φόρτος εργασίας είναι δοσοληψίες (OLTP) και το πρόβλημα είναι πώς το σύστημα θα υποδεχθεί, θα φιλτράρει, θα διαχειριστεί και θα αποθηκεύσει τα δεδομένα που ρέουν με πολύ γρήγορο ρυθμό. Τα συμβατικά συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων αδυνατούν να καλύψουν τις ανάγκες, αφού αναπόφευκτα επεξεργάζονται πολύ μεγάλη επιβάρυνση στα δεδομένα για λόγους κλειδώματος, logging, και latching σε πολυνηματικές εφαρμογές. Ο ρυθμός ανανέωσης των υπαρχόντων, σχετίζεται με το χρόνο που χρειάζεται ώστε να αντλήσουμε πληροφορία από τα εισερχόμενα δεδομένα (stream analysis / mining). Αξίζει να σημειωθεί ότι δεν είναι αρκετό να μπορούμε να αναλύσουμε τα δεδομένα και να αντλήσουμε πληροφορία σε πραγματικό χρόνο, αλλά είναι εξίσου σημαντικό να εκτελούμε και λειτουργίες που ενεργοποιούνται από αυτά, σε πραγματικό χρόνο. Μια εφαρμογή παρακολούθησης των τιμών του χρηματιστηρίου θα ήταν επιθυμητό όχι απλώς να καταγράφει τις διακυμάνσεις των τιμών αλλά και να παρέχει τη δυνατότητα αγοραπωλησίας μιας μετοχής.

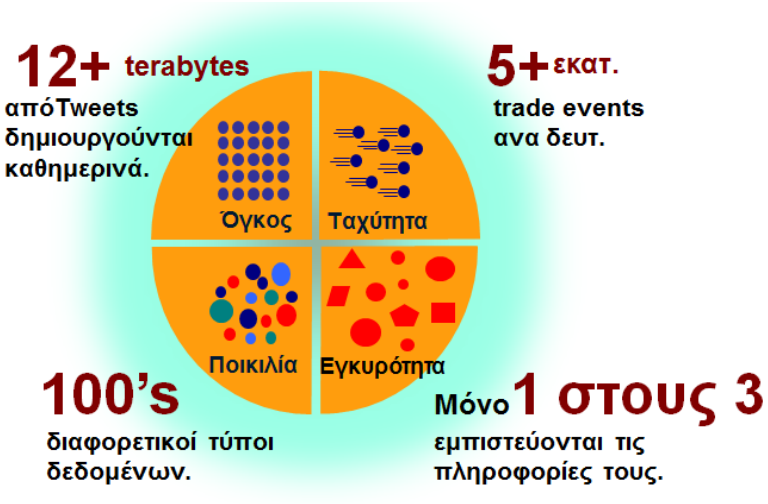
Ποικιλομορφία (Variety)

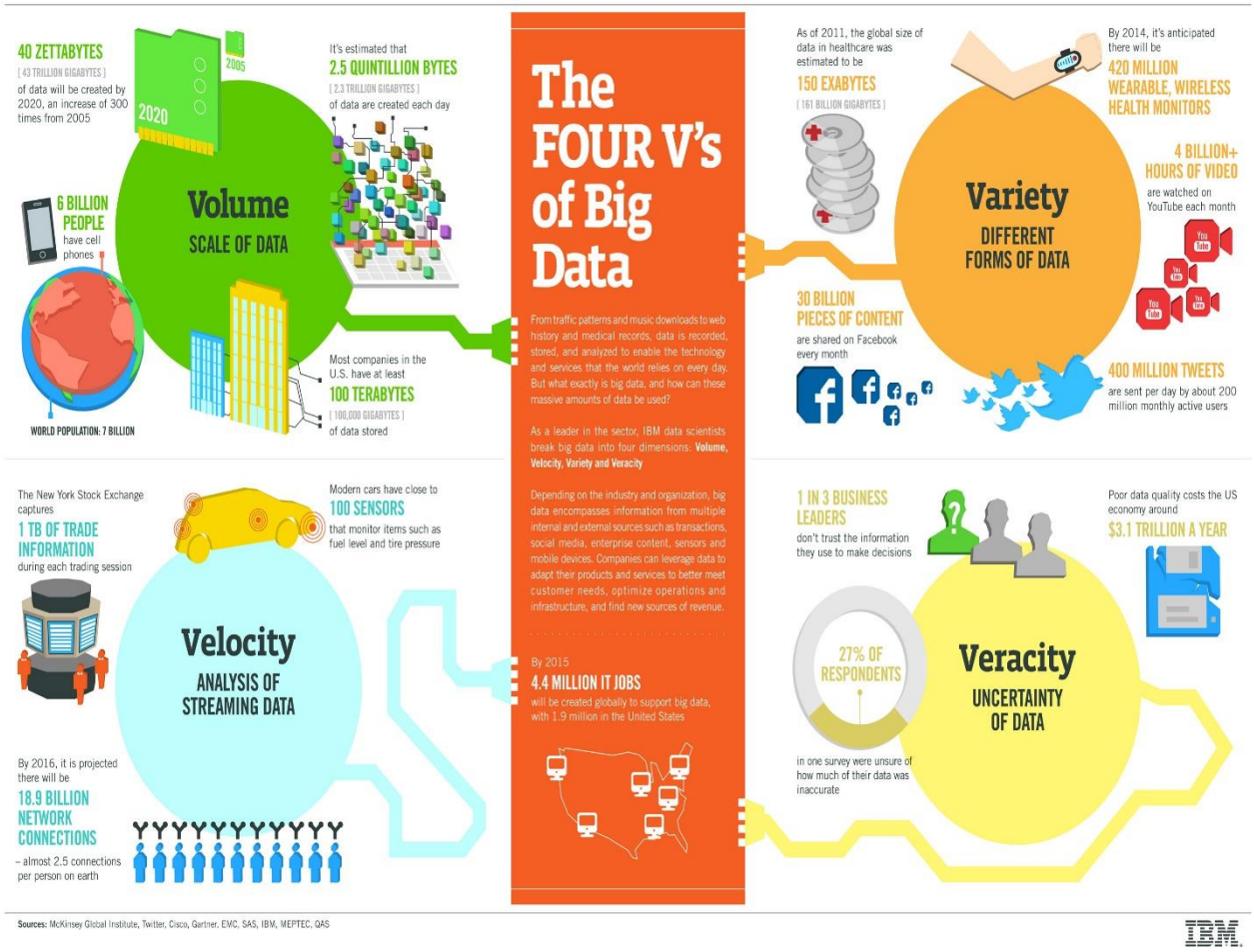
Το τρίτο γνώρισμα των μεγάλων δεδομένων είναι η ποικιλομορφία. Σήμερα καλούμαστε να αποθηκεύσουμε, να συνδυάσουμε και να επεξεργαστούμε δεδομένα από πολλές διαφορετικές πηγές π.χ κινητά δίκτυα επικοινωνιών, tablets, κάμερες, κοινωνικά δίκτυα, εταιρείες εμπορίας πληροφοριών (data brokers) κτλ. Τα δεδομένα αυτά εισρέουν σε οποιαδήποτε μορφή, δομημένα δεδομένα, αριθμητικά δεδομένα αποθηκευμένα σε παραδοσιακές βάσεις, πληροφορίες που δημιουργούνται από εμπορικές εφαρμογές, αδόμητα έγγραφα κειμένου, email, video, ήχου, δεδομένα χρηματιστηριακών συναλλαγών και εμπορικών συναλλαγών. Έτσι, έχουμε να κάνουμε όχι μόνο με διαφορετικούς τύπους δεδομένων, αλλά και με διαφορετική δομή μεταξύ ίδιων τύπων. Γεννάται έτσι η απαίτηση να ενσωματωθούν δεδομένα με αυστηρή δόμηση (structured), ημιδομημένα (semi-structured) και αδόμητα (unstructured). Ακολούθως, ακόμα και αν οι πηγές μας χρησιμοποιούν αυστηρή δόμηση των δεδομένων, πιθανόν να είναι ετερογενή, δηλαδή η δόμηση της μιας να μην είναι συμβατή με κάποια άλλη, να χρησιμοποιούν διαφορετική σημασιολογία κλπ. Συνεπώς προκύπτει ότι τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων που απαιτούν αυστηρή δομή στα δεδομένα τους, δεν μπορούν να ανταποκριθούν σε αυτές τις νέες προκλήσεις.

Εγκυρότητα (Veracity)

Πρόσφατα, πολλοί ερευνητές τονίζουν ολοένα και περισσότερο μια ακόμα πτυχή προσθέτοντας ένα τέταρτο «v»: την έννοια veracity (πιστότητα, εγκυρότητα). Η προσέγγιση αυτή δεν σχετίζεται τόσο με τα ιδιαίτερα χαρακτηριστικά των μεγάλων δεδομένων αλλά

κυρίως με το πώς πρέπει να πραγματοποιείται η χρήση τους έτσι ώστε να επιτυγχάνεται ο αναγκαίος βαθμός εμπιστοσύνης για την αξιοπιστία των δεδομένων. Το να διαθέτεις πολλά δεδομένα σε διαφορετικούς όγκους ρέοντα σε υψηλές ταχύτητες δεν ωφελεί σε κάτι αν τα δεδομένα δεν είναι ακριβή. Εσφαλμένα δεδομένα μπορούν να προκαλέσουν πολλά προβλήματα και στους οργανισμούς όσο και στους καταναλωτές. Συνεπώς, οι οργανισμοί πρέπει να διασφαλίζουν ότι τα δεδομένα είναι σωστά, καθώς και ότι οι αναλύσεις που πραγματοποιήθηκαν στα δεδομένα είναι σωστές. Ειδικά σε αυτοματοποιημένες διαδικασίες λήψης αποφάσεων, όπου κανένας άνθρωπος δεν συμμετέχει πια, πρέπει να υπάρχει σιγουριά ότι τόσο τα στοιχεία και οι αναλύσεις είναι σωστές. Η κύρια υπόσχεση των μεγάλων δεδομένων είναι η λήψη καλύτερων αποφάσεων βασιζόμενοι σε δεδομένα. Η ιδέα φαίνεται ελκυστική αλλά υπάρχει μια προειδοποίηση: είναι τα στοιχεία αρκετά αξιόπιστα για να στηρίξουν τις αποφάσεις μας πάνω σε αυτά; σε ποιο βαθμό μπορούμε να εμπιστευθούμε τα δεδομένα; Πολλές φορές το μεγαλύτερο ποσοστό του χρόνου ενός έργου αφορά τη διαλογή και καθαρισμό των δεδομένων. Το θέμα είναι ότι τα περισσότερα δεδομένα στην εποχή των μεγάλων δεδομένων είναι αβέβαια.





Σχήμα 2: Τα μεγάλα δεδομένα επεκτείνονται σε 4 μέτωπα (http//www01.ibm.com/software/data/bigdata)

<u>Αξία (Value)</u>

Επίσης, υπάρχει μια ακόμα παράμετρος η οποία μπορεί να ληφθεί υπόψη κατά την εξέταση μεγάλων δεδομένων, ένα πέμπτο V: Η Αξία (Value). Η πρόσβαση σε μεγάλα δεδομένα αν δε μπορούμε να τη μετατρέψουμε σε αξία είναι άχρηστη. Για αυτό πολλοί υποστηρίζουν ότι η αξία είναι το πιο σημαντικό V των μεγάλων δεδομένων. Οι οργανισμοί καλούνται να επιλέξουν την πιο αποτελεσματική από πλευράς κόστους λύση με στόχο την αξιοποίηση της πληροφορίας που θα οδηγεί στην έγκαιρη και όσο το δυνατό πιο σωστή λήψη αποφάσεων, δίνοντας το μεγαλύτερη δυνατόν αξία στην επιχείρηση.

Σχήμα 3 : Τα 5V των Μεγάλων Δεδομένων. Προσθέτοντας και την Αξία

2.6 ΤΕΧΝΟΛΟΓΙΕΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Η αναγκαιότητα διαχείρισης των δεδομένων σε εφαρμογές μεγάλων δεδομένων επέβαλε τη δημιουργία μίας νέας γενιάς συστημάτων, μοντέλων και προγραμματιστικών εργαλείων όπως: Map Reduce, Hadoop και οικοσύστημα αυτού, NoSQL, κ.α., τεχνολογίες που επιτρέπουν την παράλληλη επεξεργασία δεδομένων σε μεγάλη κλίμακα και με ανεκτικό στα σφάλματα τρόπο.

MapReduce

Το ***MapReduce*** αποτελεί το σπουδαιότερο εργαλείο που έχει αναπτυχθεί για την ανάλυση μεγάλων δεδομένων. Είναι ένα προγραμματιστικό μοντέλο, μαζί με τη σχετική υλοποίηση για δημιουργία και επεξεργασία πολύ μεγάλων συνόλων δεδομένων. Αναπτύχθηκε στη Google από τους Jeffrey Dean και Sanjay Ghemawat το 2004 (Chemawat, 2004)4. Έναυσμα για τη δημιουργία του ήταν το μεγάλο πλήθος υπολογισμών που εκτελούνταν ημερησίως στη Google σε πολύ μεγάλο όγκο εισερχόμενων δεδομένων. Εξαιτίας του τεράστιου αριθμού χρηστών, ο όγκος των εισερχόμενων δεδομένων επίτασσε τη χρήση κατανεμημένων συστημάτων με εκατοντάδες ή και χιλιάδες υπολογιστές ώστε να είναι εφικτό η επεξεργασία να ολοκληρωθεί μέσα σε λογικά χρονικά πλαίσια.

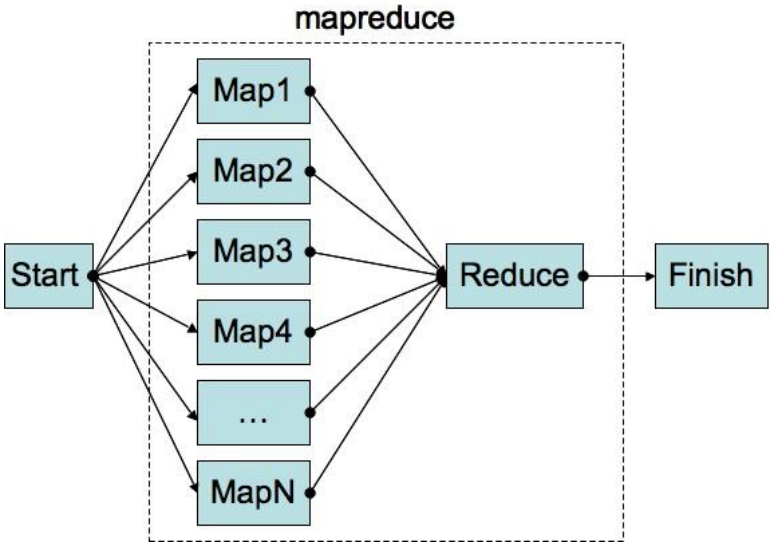
Το MapReduce είναι ένα απλό αλλά συγχρόνως πολύ δυνατό πλαίσιο το οποίο κάνει χρήση ενός αλγορίθμου ο οποίος παραλληλοποιεί και κατανέμει το σύνολο του όγκου που πρόκειται να επεξεργαστεί, μοιράζοντας κομμάτια του σε πολλούς υπολογιστές για επεξεργασία. Σε πρώτη φάση μοιράζει το σύνολο του όγκου των εργασιών σε πολλαπλούς υπολογιστές, οι οποίοι εκτελούν τα κομμάτια που τους αναθέτουν ταυτόχρονα (φάση map) και ακολούθως όλα τα αποτελέσματα συγκεντρώνονται και αναλύονται συνολικά πριν επιστραφούν (φάση reduce). Το MapReduce ουσιαστικά επιτρέπει στον προγραμματιστή να

[4] Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: simplified data processing on large clusters. In Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6 (OSDI'04), Vol. 6. USENIX Association, Berkeley, CA, USA, 10-10.

εκτελεί τις απαιτούμενες εργασίες γράφοντας 2 συναρτήσεις: τη συνάρτηση map και τη συνάρτηση reduce. Η επίλυση των προβλημάτων υλοποιείται σε 2 στάδια (Ανδρεας, 2012):

Η συνάρτηση map δέχεται σαν είσοδο ένα ζεύγος κλειδί-τιμή και παράγει σαν έξοδο ένα ζεύγος κλειδί-τιμή. Η έξοδος της συνάρτησης map, ταξινομημένη με βάση το κλειδί, είναι η είσοδος της συνάρτησης reduce. Η συνάρτηση reduce εκτελείται μετά την συνάρτηση map.
Η συνάρτηση reduce παίρνει σαν είσοδο την έξοδο της συνάρτησης map στην μορφή κλειδί-ενδιάμεσες τιμές και την επεξεργάζεται. Συνήθως για κάθε κλειδί έχουμε μία τιμή στην έξοδο.

Για την επίλυση κάποιου προβλήματος με το Map Reduce, ο προγραμματιστής πρέπει να υλοποιήσει τουλάχιστον την συνάρτηση map. Κάποιες απλές εργασίες μπορούν να υλοποιηθούν μόνο με την χρήση της συνάρτησης map.

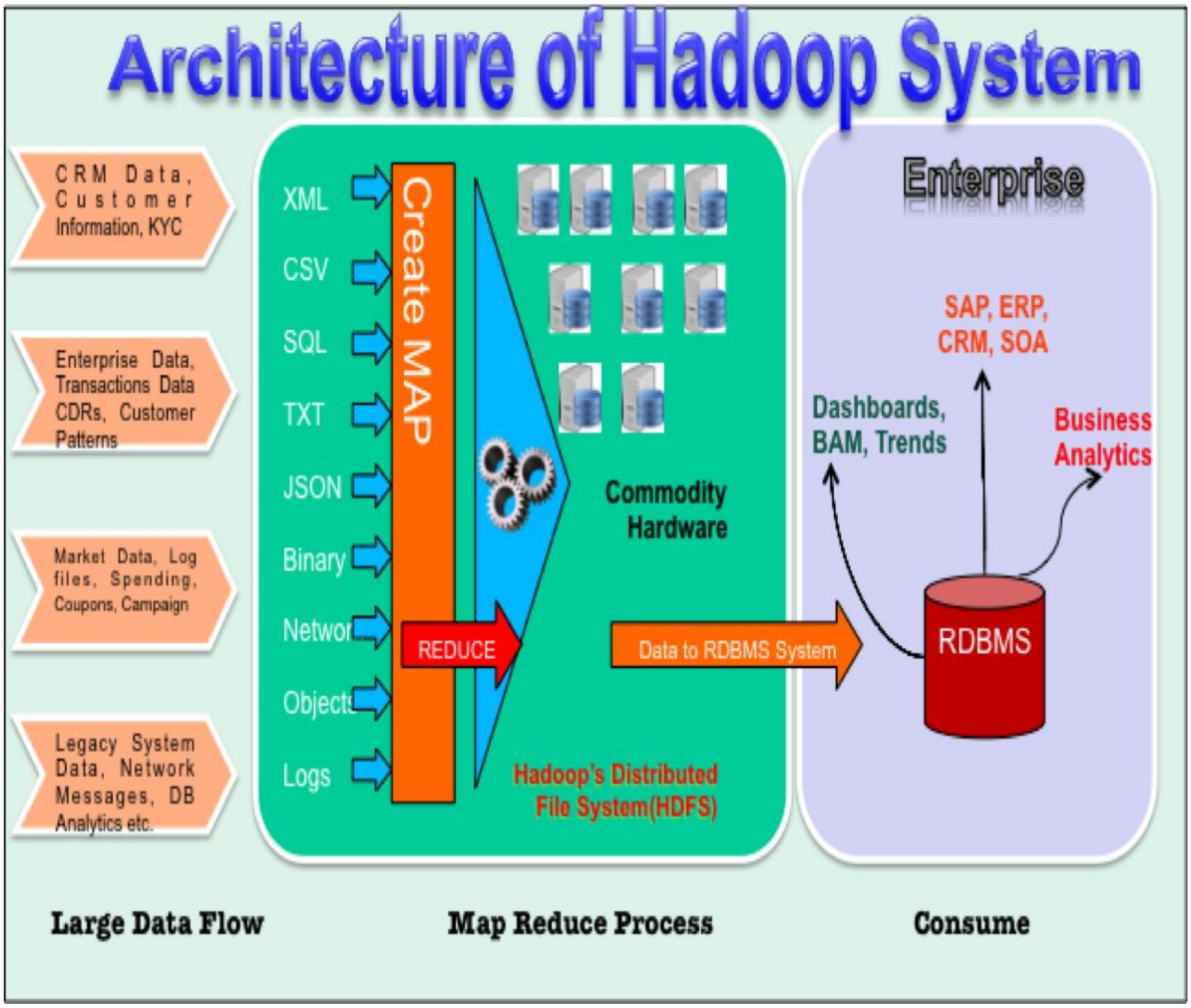


Σχήμα 4 : Φάσεις εκτέλεσης MapReduce model

Hadoop

Το Hadoop είναι ένα λογισμικό ανοιχτού κώδικα που υποστηρίζει κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων (petabytes) και παρέχει μια υλοποίηση του MapReduce. Το Hadoop βασίστηκε στο Google Map Reduce framework και το Google File System (GFS). Είναι ένα έργο του Apache Software Foundation που αναπτύσσετε και χρησιμοποιείτε από ανθρώπους από όλο τον κόσμο και κυρίως την Yahoo!. Σήμερα είναι η πιο διαδεδομένη υλοποίηση του MapReduce και χρησιμοποιείται για διδακτικούς σκοπούς σε αρκετά πανεπιστήμια του κόσμου, αλλά και σε μεγάλους οργανισμούς ανά το παγκόσμιο για την επεξεργασία μεγάλων δεδομένων εισόδου. Κάποιοι από τους οργανισμούς, που διατηρούν clusters για εκτέλεση Hadoop εργασιών είναι: Yahoo!, Amazon, AOL, Alibaba, Cornell University Web Lab, ETH Zurich Systems Group, Facebook, Google, IBM, New York Times κ.α.

Το Hadoop κατανέμει τα δεδομένα και την ανάλυσή τους σε ομάδες υπολογιστών (Clusters) ώστε να επεξεργαστούν ταυτόχρονα τα δεδομένα, εξοικονομώντας έτσι χρόνο και πόρους. Πιο συγκεκριμένα, προωθεί τα δεδομένα και το πρόβλημα στον υπολογιστή master της ομάδας, και αυτός στη συνέχεια θα κατακερματίσει το πρόβλημα σε μικρότερα προβλήματα και κάθε νέο πρόβλημα θα το προωθήσει με τη σειρά του σε κάθε ένα από τους υπόλοιπους υπολογιστές της ομάδας. Κάθε υπολογιστής της ομάδας θα επιλύσει το δικό του μικρότερο πρόβλημα και θα επιστρέψει τη λύση στον master υπολογιστή ο οποίος θα συνδυάσει τις λύσεις των υποπροβλημάτων για να βρει τη λύση στο αρχικό πρόβλημα. Το Hadoop στηρίζεται σε αυτή την διαδικασία ενώ επιπλέον προτέρημα είναι ότι επιτυγχάνει να ανακτήσει δεδομένα σε περίπτωση που ένας υπολογιστής της ομάδας πάθει ζημιά και να μεταβιβάσει το υποπρόβλημα σε άλλον υπολογιστή.



Σχήμα 5: Παράδειγμα αρχιτεκτονικής ενός Hadoop Cluster (http://technologist-work.com/wp-content/uploads/2014/09/Hadoop.png)

Πλεονεκτήματα του Hadoop

Τα σημαντικότερα πλεονεκτήματα του Hadoop μπορούν να συνοψιστούν ως εξής:

Επεκτασιμότητα: Δυνατότητα αξιόπιστης αποθήκευσης και επεξεργασίας μέχρι και petabytes δεδομένων

Οικονομία Πόρων: Κατανομή δεδομένων και επεξεργασίας σε ομάδες υπολογιστών που αποτελούνται από έως και χιλιάδες κοινούς υπολογιστές.

Αποδοτικότητα: Με την κατανομή των δεδομένων, η επεξεργασία γίνεται ταυτόχρονα σε όλους τους κόμβους, παρέχοντας ταχεία εκτέλεση των εργασιών.

Αξιοπιστία (fault tolerance): Επιτυγχάνεται μέσω της αυτόματης διατήρησης πολλαπλών αντιγράφων των δεδομένων, καθώς και αυτόματης ανάθεσης των εργασιών υπολογισμού σε νέους κόμβους σε περίπτωση βλάβης.

Κριτική στο Hadoop

Παρά τα πλεονεκτήματά του και τη θόρυβο που γίνεται γύρω από αυτό, το Hadoop δεν σημειώνει εξίσου υψηλό βαθμό δημοτικότητας από όλους τους επιστήμονες δεδομένων. Στην πράξη, δεν είναι λίγοι εκείνου που το χρησιμοποίησαν και το εγκατέλειψαν. Σε έρευνα για τα εμπόδια των big data analytics, το Paradigm σημειώνει ότι το 76% των επιστημόνων δεδομένων που χρησιμοποιούσαν Hadoop, απάντησαν ότι έχει «σημαντικούς περιορισμούς»5 ().

Μία αιτία απογοήτευσης είναι το κόστος του Hadoop. Πολλοί οργανισμοί επιλέγουν το Hadoop διότι θεωρούν ότι θα εξοικονομήσουν χρήματα επειδή είναι ανοικτού κώδικα, ωστόσο διαπιστώνουν το αντίθετο. Στην πραγματικότητα αναγκάζονται να πληρώνουν για επιπλέον υπηρεσίες της Hadoop και για πρόσληψη προγραμματιστών και αναλυτών.

Οι οργανισμοί που δοκίμασαν το Hadoop και αντιμετωπίσανε προβλήματα ενδεχομένως να αποδειχθούν τα πρώτα θύματα του πρώτου κύματος της μόδας του Hadoop. Η σταδιακή ωρίμανση των μεγάλων δεδομένων και της τεχνολογίας analytics, ταυτόχρονα με καλύτερα εκπαιδευμένους χρήστες, θα καταστήσουν πιο εφικτή τη δυνατότητα εξεύρεσης καλύτερης λύσης για analytics.

2.7 ΕΦΑΡΜΟΓΕΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Καθώς η τεχνολογία των μεγάλων δεδομένων διαρκώς εξελίσσεται, ολοένα και πιο πολλοί οργανισμοί αντιλαμβάνονται την ανάγκη και ευκαιρία αξιοποίησης τους. Ορισμένα "κέρδη" από τις εφαρμογές της ανάλυσης μεγάλων δεδομένων είναι:

- Ανάλυση της συμπεριφοράς και των προτιμήσεων των χρηστών και ανάλογη προσαρμογή της πολιτικής προβολής και διαφήμισης
- Εντοπισμός μοτίβων που αφορούν στην ανταπόκριση ασθενών σε ιατρικές θεραπείες

[5] Leaving Data on the Table: Data Scientists Reveal Obstacles to Big Data Analytics, Paradigm4 Data Scientist Survey, http://www.citeworld.com/article/2462886/big-data-analytics/when-to-use-hadoop-and-when-not-to.html

- Εντοπισμός αναζητουμένων ή/και υποψηφίων εγκληματιών μέσω της ανίχνευσης υπόπτων τηλεπικοινωνιακών, αγοραστικών και μετακινήσεων συμπεριφορών
- Εφαρμογές για κινητές συσκευές: Τα τελευταία χρόνια οι κινητές συσκευές αποκτούν όλο και μεγαλύτερο μερίδιο της αγοράς για τις διαδικτυακές υπηρεσίες. Οι υπηρεσίες αυτές για να ανταποκριθούν κατάλληλα, θα πρέπει να εξασφαλίσουν υψηλή διαθεσιμότητα αλλά και τη δυνατότητα για άμεση επεξεργασία μεγάλου πλήθους δεδομένων. Τα δύο αυτά χαρακτηριστικά υποδεικνύουν τα big datas ως την καλύτερη λύση.
- Εφαρμογές με υψηλό βαθμό παραλληλοποίησης: Η σημερινή έκρηξη δεδομένων έχει ως συνέπεια την ανάγκη για γρήγορη επεξεργασία τεράστιου όγκου πληροφορίας. Αυτό με τις συμβατικές τεχνολογίες είναι αδύνατον να επιτευχθεί. Στο clouding ωστόσο, με την ύπαρξη εικονικά απεριόριστου hardware (on demand), μπορούμε να μοιράσουμε τα δεδομένα μας σε εκατοντάδες διαφορετικούς υπολογιστές για να επιτύχουμε γρηγορότερη επεξεργασία τους. Μάλιστα, οι πάροχοι διαθέτουν έτοιμα περιβάλλοντα (MapReduce, Hadoop) για τη διευκόλυνση αυτών των εφαρμογών.
- Επιχειρησιακές εφαρμογές (business applications): Μεγάλα συστήματα (ERP, CRM κλπ) που χρησιμοποιούνται για το στρατηγικό σχεδιασμό των επιχειρήσεων, είναι συνήθως πολύ απαιτητικά σε υπολογιστική ισχύ. Η λύση του cloud λοιπόν φαντάζει μονόδρομος, από τη στιγμή που το μέγεθος των δεδομένων προς επεξεργασία για τις μεγάλες επιχειρήσεις αυξάνεται εκθετικά.

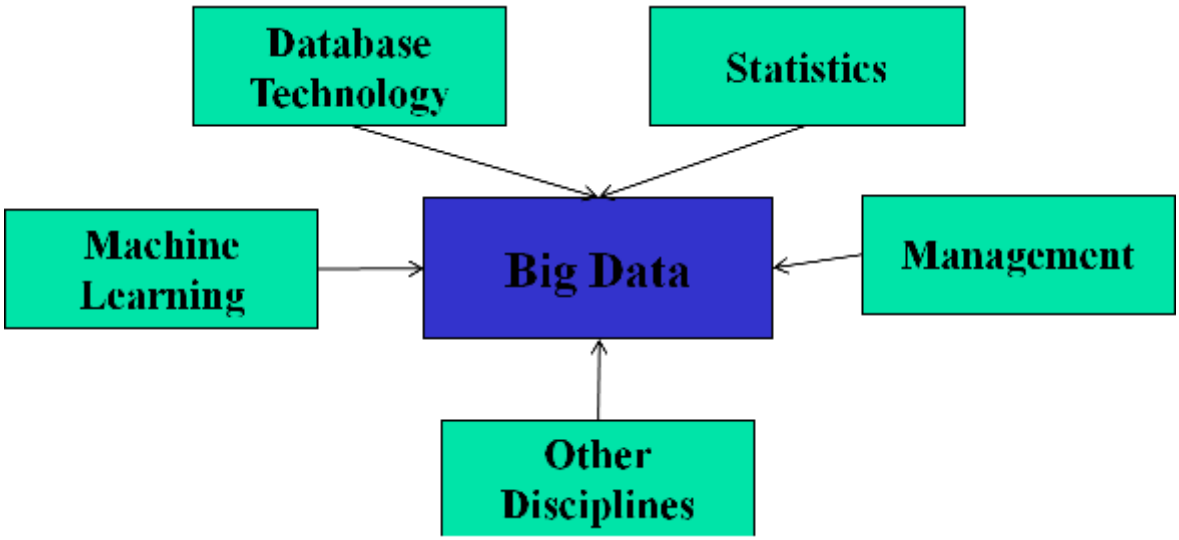
Πολλές είναι οι εταιρείες οι οποίες κινηθήκαν με ταχύτητα και κατάφεραν να επιτύχουν επιχειρηματικούς στόχους μέσα από την ανάλυση μεγάλων δεδομένων. Εταιρίες που τα εκμεταλλεύτηκαν άμεσα, όπως η T-Mobile πάροχος υπηρεσιών κινητής τηλεφωνίας και θυγατρική της Deutche Telecom, πέτυχε τη μείωση κατά 50% του λειτουργικού κόστους χάρη στα μεγάλα δεδομένα. Πολλά παραδείγματα εταιρειών που με χρήση μεγάλων δεδομένων βελτίωσαν υπηρεσίες μεταφορών, γεωργίας, χρηματοοικονομικών και παιδείας και μπορούν να χωριστούν σε διάφορες κατηγορίες ανάλογα με την περίπτωση.

Μέσω των analytics μεγάλων δεδομένων είναι εφικτή η παρακολούθηση της εξέλιξης των πωλήσεών. Η T-mobile, περιόρισε το κόστος αναλύοντας μια σειρά πολλαπλών δεικτών δεδομένων προκειμένου να επιτύχουν τη βελτίωση της εξυπηρέτησης πελατών και την επίτευξη των στόχων τους.

- Δεδομένα που αφορούν τη συμπεριφορά των πελατών αποτελούν πολύτιμη πηγή πληροφορίας προς εντοπισμό των κινήτρων που διαμορφώνουν τις επιλογές των καταναλωτών σχετικά με ένα προϊόν. Η Time Warner Cable, καταγράφει τη συμπεριφορά των πελατών, τη δραστηριότητά τους, την χρονική κατανομή της

κατανάλωσης μέσα σε μια ημέρα προκειμένου να προσαρμόσουν την πολιτική τους σχετικά.

Στη συνέχεια περιγράφεται ο τρόπος αξιοποίησης των μεγάλων δεδομένων σε μια σειρά από σημαντικούς τομείς της κοινωνικής και οικονομικής δραστηριότητας.



Σχήμα 6 :Εφαρμογές big datas
(http://analytics.dmst.aueb.gr/sites/all/themes/bluez/images/periexomeno/periexomeno2.png)

Τηλεπικοινωνίες

Στα τηλεφωνικά κέντρα συλλέγονται πολύ μεγάλες ποσότητες αδόμητων και δομημένων δεδομένων. Η χαρτογράφηση και ταξινόμηση των κλήσεων παρέχει τη δυνατότητα εντοπισμού σφαλμάτων και αδυναμιών στις σχετικές υποδομές (ESRL, 2011)

Εμπόριο

Η eBay, μία από τις μεγαλύτερες ηλεκτρονικές πλατφόρμες δημοπρασιών στον κόσμο, καταγράφει συναλλαγές με περισσότερους από 108 εκατ. πελάτες ετησίως ενώ εισπράττει πάνω από 250 εκατ αιτήματα στον ιστοχώρο της ημερησίως. Επίσης, στο εμπόριο ηλεκτρονικών συσκευών, εκτιμάται ότι πωλείται ένα κινητό τηλέφωνο κάθε 5 δευτερόλεπτα (Berkeley, 2012). Είναι εμφανώς λοιπόν ο όγκος πληροφορίας που αποθηκεύεται τόσο για τους πελάτες και για τα προϊόντα και η δυνατότητα εξόρυξης γνώσης αναφορικά με τις συνήθειες και τάσεις και η διαμόρφωση σχετικών πολιτικών. Αλλά ακόμα και στο λιανεμπόριο γίνεται αξιοποίηση μεγάλων δεδομένων μελετώντας πληροφορίες από MME για τον εντοπισμό του βέλτιστου σημείου εγκατάσταση ενός καταστήματος.

Αντιμετώπιση καταστροφών

Μέσα από τη συγκέντρωση πληροφορίας που μπορεί να προέρχεται είτε από MME είτε από κοινό δίνεται η δυνατότητα χαρτογράφησης του τόπου όπου λαμβάνει χώρα μια καταστροφή, αξιολόγησης της σοβαρότητάς της και χάραξη της άριστης διαδρομής για την

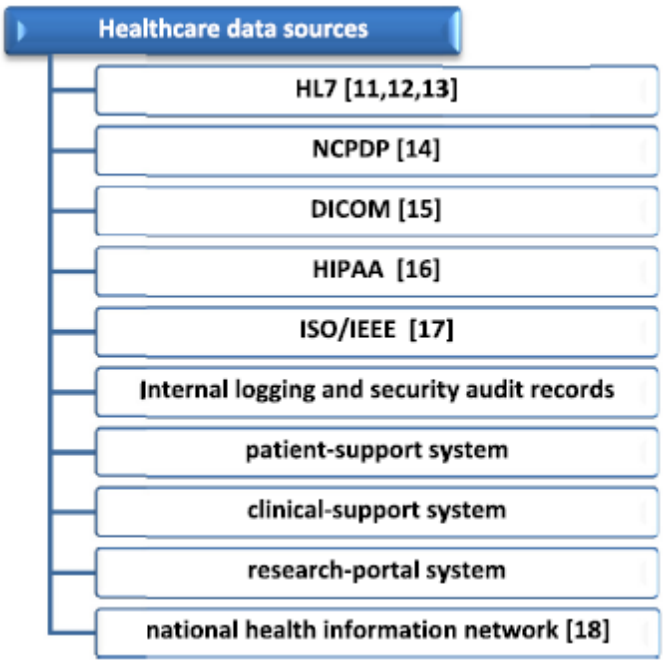
τάχιστη άφιξη των σχετικών υπηρεσιών. Η γρήγορη πρόγνωση του τυφώνα Irene στην Φλόριντα το 2011 ελαχιστοποίησε τις συνέπειες αφού έδωσε το χρόνο για τη λήψη όλων των αναγκαίων μέτρων και βασίστηκε στην ανάλυση μεγάλων γεωχωρικών δεδομένων

Φυσικοί πόροι

Ο κλάδος του πετρελαίου θεωρείται από τους πρώτους όπου άρχισε να ασχολείται με τα μεγάλα δεδομένα. Πετρελαϊκές εταιρείες και κυβερνήσεις κάνουν χρήση και ανάλυση τεράστιων ποσοτήτων δεδομένων που είναι διαθέσιμα σχετικά με σεισμική δραστηριότητα σε όλη την υφήλιο με σκοπό την εξερεύνηση και εξόρυξη πετρελαίου.

Υγεία

Στις μέρες μας, τα δεδομένα που αφορούν ασθενείς δημιουργούνται με εκθετικό ρυθμό. Οι πληροφορία αυτές ωστόσο έχουν διαφορετικές μορφές και πρότυπα. Σύμφωνα με τους Liu και Park (Big Data as an e-health service, 2014), υπάρχουν διάφορες πηγές δεδομένων, όπως φαίνεται στο παρακάτω σχήμα.



Σχήμα 7 : Πηγές Δεδομένων στον Τομέα της Υγείας

Όπως φαίνεται στο Σχ. 2, τεράστιος όγκος πληροφοριών παράγονται σε διάφορες μορφές με υψηλή ταχύτητα.. Στα δεδομένα που σχετίζονται με προβλήματα υγείας παρουσιάζονται όλα τα V και, επομένως, είναι σίγουρο ότι θα χρησιμοποιηθούν λύσεις μεγάλων δεδομένων για την επίλυσή τους. Παρ 'όλα αυτά, σύμφωνα με το Liu και Park, οι υπάρχουσες τεχνολογίες μεγάλων δεδομένων δεν αντιμετωπίζουν επαρκώς όλο το φάσμα των προβλημάτων υγείας, γι' αυτό είναι απαραίτητο να προσαρμόζονται ανάλογα με τους σκοπούς. Επίσης, μειώνεται το κόστος χρησιμοποιώντας ανάλυση μεγάλων δεδομένων στον τομέα της υγείας (Uma Srinivasan, 2013). Οι Chawla και Davis (Nitesh V. Chawla, 2013),

υποστηρίζουν ότι με μία ανάλυση μεγάλων δεδομένων θα ήταν εφικτή μια ασθενοκετρική προσέγγιση του τομέα Υγείας.

Μεταφορές

Η εταιρεία διεθνών ταχυμεταφορών UPS, άρχισε να καταγράφει και να τις κινήσεις πακέτων και συναλλαγών από τις αρχές της δεκαετίας 1980. Σήμερα συγκεντρώνει δεδομένα για 16,3 εκατ. πακέτα ημερησίως, για 8,8 εκατ πελάτες, με μέσω όρο 39,6 εκατ αιτήματα παρακολούθησης πακέτου καθημερινά. Το μεγαλύτερο τμήμα των μεγάλων δεδομένων που διαθέτει προέρχονται από τηλεματικούς αισθητήρες τοποθετημένους στα οχήματά της. Τα δεδομένα αξιοποιούνται τόσο για την καθημερινή εποπτεία και μέτρηση της αποδοτικότητας αλλά και για τη διαμόρφωση της βέλτιστης δομής των δρομολογίων. Συνέπεια, της εν λόγω πρακτικής ήταη η εξοικονόμηση πάνω από 8.4 εκατ γαλονιών καυσίμων το 2011, και η μείωση κατά 85 εκατ μίλια τα καθημερινά δρομολόγια. Η UPS εκτιμά ότι η μείωση ενός μιλίου ανά οδηγό την ημέρα, εξοικονομεί την κόστος 30.000.000 δολαρίων.

Διαδίκτυο

Ιστότοποι όπως το facebook και το twitter συγκεντρώνουν πάνω από 25 και 12 terabytes δεδομένων αντίστοιχα. Η Google μέσω των διάφορων εφαρμογών της (mail, google drive, google earth κ.α) συγκεντρώνει δεδομένα όγκου πλέον των 80 terabytes ημερησίως. Η ανάλυση των δεδομένων των χρηστών τους είναι ο οδηγός της διαμόρφωσης της στρατηγικής τους στόχευσης.

Κεφάλαιο 3ο

3.0 ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Λόγω του τεράστιου όγκου διαθέσιμων δεδομένων χρειαζόμαστε μεθόδους για να τα αναλύσουμε. Η τεχνολογία του data mining είναι μια περιοχή η οποία περιλαμβάνει τεχνικές επεξεργασίας και ανάλυσης μεγάλων βάσεων δεδομένων. Ο στόχος αυτών των τεχνικών είναι η ανακάλυψη νέων προτύπων μεταξύ των δεδομένων και η εξαγωγή χρήσιμων πληροφοριών. Στην παρούσα πτυχιακή εργασία μελετήθηκαν και παρουσιάζονται ορισμένες από αυτές τις τεχνικές.

3.1 ΜΕΘΟΔΟΛΟΓΙΑ ΔΗΜΙΟΥΡΓΙΑΣ ΜΟΝΤΕΛΩΝ

Η διαδικασία δημιουργίας μοντέλων αναπτύχθηκε και τελειοποιήθηκε από πολλούς επαγγελματίες στη διάρκεια πολλών ετών. Εδώ παρουσιάζεται μια απλή και αποδεδειγμένη προσέγγιση δημιουργίας επιτυχημένων και επικερδών μοντέλων.

Προετοιμάστε τα δεδομένα. Αυτό το βήμα πρέπει να ολοκληρωθεί πριν από οποιαδήποτε ουσιαστική διερεύνηση ή ανάλυση λάβει χώρα. Σε μεγάλους οργανισμούς, αυτό γίνεται από μια ξεχωριστή ομάδα και πιθανόν διαφορετική επιχειρηματική μονάδα. Ανεξαρτήτως του πως τα δεδομένα έχουν προετοιμαστεί, είτε μέσω επίσημου αιτήματος με συνοδευτικό προδιαγραφών, είτε εάν έχετε λάβει άδεια να απευθύνετε απευθείας το ερώτημα από την αποθήκη δεδομένων της επιχείρησης (EDW), το βήμα αυτό είναι ουσιώδες για ένα επιτυχές μοντέλο. Η επένδυση στην κατανόηση της διαδικασίας προετοιμασίας των δεδομένων για την επιχείρησή σας συχνά αποβαίνει επωφελής σε μακροπρόθεσμο ορίζοντα. Καθώς χρειάζεται πρόσβαση σε ολοένα και μεγαλύτερα και πιο αναλυτικά δεδομένα, η γνώση του ποια δεδομένα υπάρχουν και πως μπορούν να συνδυαστούν με άλλες πηγές δεδομένων θα παράσχει γνώση που δεν ήταν δυνατή μόλις λίγα χρόνια πριν.

Εάν η επιχείρηση πληροφορικής δεν διατηρεί περισσότερα δεδομένα για μεγαλύτερο χρονικό διάστημα και σε τελειότερα επίπεδα τότε βρίσκεστε πίσω από την τάση της εποχής και κινδυνεύετε να καταστείτε άσχετοι στην αγορά σας.

Εκτέλεση διερευνητικής ανάλυσης δεδομένων.

Αυτό είναι το βήμα όπου κατανοούνται τα δεδομένα και ξεκινά να αποκτά κάποιος διαίσθηση των σχέσεων μεταξύ των μεταβλητών. Αυτή η διερεύνηση γίνεται καλύτερα με έναν ειδικό στον τομέα εάν δεν υπάρχει εμπειρία. Φανταστικοί ανθρακωρύχοι δεδομένων θα ανακαλύψουν τις σχέσεις και τις τάσεις μέσω μιας πολύπλοκης εξερεύνησης. Με μεγάλο ενθουσιασμό θα παρουσιάσουν τα ευρήματα τους στον ειδικό του τομέα ο οποίος ευγενικά θα απαντήσει πως αυτά τα πράγματα είναι γνωστά εδώ και καιρό.

Δεν μπορούμε να δώσουμε μια ολοκληρωμένη και διεξοδική ανάλυση και σύσταση κάτω απο τις σημερινές επιχειρηματικές συνθήκες χωρίς εμπειρία στον τομέα που θα συμπληρώσει την αναλυτική σας ικανότητα.

Τα εργαλεία που χρησιμοποιούνται για την διερευνητική ανάλυση δεδομένων έχουν αλλάξει τα τελευταία χρόνια. Στα τέλη της δεκαετίας του ’90 αυτό γινόταν χρησιμοποιώντας κυρίως αποτελέσματα σε πίνακες SQL ή κάποιας γλώσσας script. Έπρεπε να μελετηθεί η συσχέτιση πινάκων και στατιστικών γραφημάτων. Η δουλειά αυτή ήταν συχνά αργή και απαιτούνταν ικανότητες προγραμματισμού. Τα τελευταία χρόνια τα γραφικά εργαλεία έχουν βελτιωθεί δραματικά. Υπάρχουν πλέον πολλά εμπορικά προιόντα από την SAS την IBM και την SAP αλλά και μικρότερων προμηθευτών όπως της QlikTech και της Tableau, για να αναφέρουμε μερικές, στον χώρο της επιχειρηματικής οπτικοποίησης. Αυτά τα προιόντα είναι ικανά να φορτώσουν δεδομένα για οπτική διερεύνηση, συνήθως μέσω ενός μέσου αλληλεπίδρασης τύπου ‘browser’ και να παρέχουν μια άκρως διαδραστική εμπειρία διερεύνησης δεδομένων. Αυτή η τεχνολογία έχει αποδειχτεί πως λειτουργεί με δισεκατομμύρια παρατηρήσεις υπό την προπόθεση πως υπάρχουν επαρκείς πόροι φυσικών υλικών.

Η διερεύνηση των δεδομένων δεν ολοκληρώνεται ποτέ. Υπάρχουν πάντα περισσότεροι τρόποι να εξετάσουμε τα δεδομένα και τις αλληλεπιδράσεις και σχέσεις που πρέπει να λάβουμε υπ’όψην, έτσι ώστε να παρατηρείται η αρχή της επάρκειας και ο νόμος των φθίνουσων επιστροφών. Ο νόμος των φθίνουσων επιστροφών προέρχεται από το πεδίο των οικονομικών και ορίζει πως για κάθε μονάδα προσπάθειας (χρόνου στην περίπτωσή μας) που προσθέτουμε, αποδίδεται λιγότερο αυξημένη αξία ανά μονάδα για κάθε επιτυχή μονάδα προσπάθειας που εναποθέτουμε στο έργο μας. Σε αυτή την περίπτωση, η διορατικότητα και η γνώση που αποκτάται μεταξύ των ωρών 15 και 16 της διερεύνησης δεδομένων είναι πολύ πιθανόν λιγότερη από την διορατικότητα που αποκτήθηκε μεταξύ των ωρών 2 και 3. Η αρχή της επάρκειας επιβεβαιώνει τον νόμο των φθίνουσων επιστροφών και θέτει το όριο της παραγωγικής απώλειας. Με απλά λόγια: Μάθε πότε να σταματάς την διερεύνηση. Η μεθοδολογία ανάπτυξης λογισμικού προχώρησε στην εφαρμογή αυτής της ιδέας μέσω ευκίνητων διαδικασιών εκμάθησης ολίγων, για αρχή, και συνεχούς βελτίωσης. (Dean, 2014)

<u>Δημιουργία του πρώτου μοντέλου.</u>

Το κλειδί για αυτό το βήμα είναι να συνειδητοποιήσει κανείς εκ των προτέρων πως η επιτυχής διαδικασία δημιουργίας μοντέλων θα περιλαμβάνει πολλές επαναλήψεις. Κατά την διάρκεια κάποιων εργασιών η φράση του Thomas Edison οτι *«Δεν απέτυχα. Απλά ανακάλυψα 10000 τρόπους που δεν δουλεύουν»* θα φανεί πολύ εύστοχη. Μέχρι να δημιουργήσετε το πρώτο μοντέλο, δεν θα είστε σε θέση να αξιολογήσετε ποια θα είναι η πιθανή του επίπτωση. Δημιουργήστε το μοντέλο γρήγορα με την μέθοδο όπου αισθάνεστε πιο άνετα. (Συχνά χρησιμοποιούμαι ένα ‘δένδρο’ αποφάσεων). Η δημιουργία του πρώτου μοντέλου βοηθάει να

δομήσετε τα κριτήρια της επιτυχίας και να θέσετε κατάλληλες προσδοκίες για τους ανθρώπους που θα χρησιμοποιήσουν τις προβλέψεις του μοντέλου. Η ανθρώπινη φύση είναι αισιόδοξη. Πιστεύουμε σε μεγαλύτερες πωλήσεις προιόντων από όσο πραγματικά πουλάνε, σε εργασίες που θα ολοκληρωθούν χωρίς επιπλοκές και ούτω καθεξής. Η δημιουργία του πρώτου μοντέλου ουσιαστικά αποτελεί έναν έλεγχο πραγματικότητας για μελλοντικές αποδόσεις και προσδοκίες. Το πρώτο μοντέλο είναι, κατά κανόνα, ο πρωταθλητής (Dean, 2014)

Κατασκευή μοντέλων κατ' επανάληψη.

Αυτή είναι η φάση στην οποία πρέπει να ξοδέψει κάποιος τον περισσότερο χρόνο. Το βήμα αυτό είναι ένας βρόχος ανάδρασης όπου θα κτιστεί ένα μοντέλο (ο προκαλών) και θα το συγκρίνει με το μοντέλο 'πρωταθλητή' χρησιμοποιώντας κάποια αντικειμενικά κριτήρια που καθορίζουν το καλύτερο μοντέλο. Εάν ο 'προκαλών' είναι καλύτερος από τον 'πρωταθλητή' τότε εκτιμήστε κατά πόσον ο 'προκαλών' ικανοποιεί τους στόχους της εργασίας. Εάν οι στόχοι της εργασίας δεν ικανοποιούνται ή ο 'πρωταθλητής' δεν εκτοπίζεται τότε κτίστε ένα άλλο μοντέλο. Συχνά δεν υπάρχει κάποια βάσιμη αξιολόγηση μοντέλου για να καθορίσει πότε θα σταματήσετε, παρά ένα χρονικό περιθώριο που εξαναγκάζει την εργασία να τερματίσει. Ας υποθέσουμε πως έχετε δεσμευτεί να παράσχετε μια λίστα πελατών για μια εμπορική καμπάνια. Ο 'πρωταθλητής' έχει προθεσμία να παράσχει την λίστα των πελατών μέχρι την επόμενη Τρίτη, επομένως η δημιουργία μοντέλων θα συνεχιστεί μέχρι εκείνη την χρονική στιγμή. (Dean, 2014)

3.2 sEMMA

Η sEMMA είναι μια μεθοδολογία εξόρυξης δεδομένων, δημιουργία της SAS, που εστιάζει στην λογική οργάνωση της αναπτυξιακής φάσης του μοντέλου των εργασιών εξόρυξης δεδομένων. Περιγράφει την διαδικασία που κάποιος πρέπει να ακολουθήσει για να αποσπάσει βαθειά γνώση των δεδομένων του. Το ακρώνυμο sEMMA – δειγματισμός, μετατροπή, μοντελοποίηση, πρόσβαση – αναφέρεται στην βασική διαδικασία διεξαγωγής εξόρυξης δεδομένων. Ξεκινώντας με ενα στατιστικά αντιπροσωπευτικό δείγμα των δεδομένων, η sEMMA καθιστά εύκολη την διερευνητική εφαρμογή στατιστικών και οπτικών τεχνικών, την επιλογή και μετατροπή των πιο σημαντικών μεταβλητών πρόβλεψης, μοντελοποίησης των μεταβλητών και πρόβλεψης αποτελεσμάτων καθώς και επιβεβαίωσης της ακρίβειας του μοντέλου.

Προτού εξετάσουμε κάθε φάση της sEMMA, επιτρέψτε μας να αναφέρουμε μια συχνή παρανόηση. Η συχνή παρανόηση είναι να αναφερόμαστε στην sEMMA ως μια μεθοδολογία εξόρυξης δεδομένων. Η sEMMA δεν είναι κάτι τέτοιο, είναι μια λογική οργάνωση της

λειτουργικής εργαλειοθήκης της εταιρίας εξόρυξης SAS που εκτελεί τις βασικές εργασίες για εξόρυξη δεδομένων. Η εταιρία εξόρυξης μπορεί να χρησιμοποιηθεί σαν μέρος οποιασδήποτε επαναληπτικής μεθοδολογίας εξόρυξης δεδομένων που υιοθετείται από το πελάτη. Φυσιολογικά, βήματα όπως η διαμόρφωση μιας ορθά καθορισμένης επιχείρησης ή διερεύνησης προβλήματος και συναρμολόγηση ποιοτικών και αντιπροσωπευτικών πηγών δεδομένων, είναι κρίσιμης σημασίας για την συνολική επιτυχία οποιασδήποτε εργασίας εξόρυξης δεδομένων. Η sEMMA εστιάζει στις αναπτυξιακές πτυχές του μοντέλου της εξόρυξης δεδομένων.

Δειγματισμός (προαιρετικά)

Τα δεδομένα σας εξάγοντας τμήμα ενός μεγάλου συνόλου δεδομένων. Αυτό πρέπει να είναι τόσο μεγάλο ώστε να περιλαμβάνει όλες τις σημαντικές πληροφορίες αλλά και ταυτόχρονα τόσο μικρό ώστε να χειρίζεται γρήγορα. Για βέλτιστο κόστος και απόδοση, το ινστιτούτο SAS συνηγορεί προς μια στρατηγική δειγματισμού που εφαρμόζει ένα αξιόπιστο και στατιστικά αντιπροσωπευτικό δείγμα μεγάλων και λεπτομερών πηγών δεδομένων. Η εξόρυξη ενός αντιπροσωπευτικού δείγματος αντί ολόκληρου του όγκου μειώνει τον χρόνο επεξεργασίας που απαιτείται για να αποκτηθούν κρίσιμες επιχειρηματικές πληροφορίες. Εάν κάποια γενικευμένα μοτίβα εμφανιστούν στα δεδομένα ως ολότητα, αυτά θα είναι ανιχνεύσιμα σε ένα αντιπροσωπευτικό δείγμα. Εάν κάτι είναι τόσο μικρό που δεν αντιπροσωπεύεται στο δείγμα και ταυτόχρονα τόσο σημαντικό που επηρεάζει την ευρύτερη εικόνα, μπορεί να ανακαλυφθεί χρησιμοποιώντας μεθόδους περίληψης. Επίσης συνηγορούμε στη δημιουργία διαχωρισμένων συνόλων δεδομένων με τον κόμβο διαχωρισμού δεδομένων:
Εξάσκηση – χρησιμοποιείται για προσαρμογή μοντέλων.
Επικύρωση – χρησιμοποιείται για την αξιολόγηση και την πρόληψη υπερπροσαρμογής.
Δοκιμή – χρησιμοποιείται για να αποκτήσουμε μια ειλικρινή αξιολόγηση του πόσο καλά ένα μοντέλο γενικεύεται.

Εξερεύνηση των δεδομένων σας αναζητώντας απρόβλεπτες τάσεις και ανωμαλίες προκειμένου να αποκτήσετε κατανόηση και ιδέες.

Η εξερεύνηση βοηθά στην τελειοποίηση της διαδικασίας ανεύρεσης. Εάν η οπτική εξερεύνηση δεν αποκαλύπτει ξεκάθαρες τάσεις μπορείτε να εξερευνήσετε τα δεδομένα μέσω στατιστικών τεχνικών συμπεριλαμβανομένων αναλύσεων παραγόντων, αντιστοιχίας και ομαδοποίησης. Για παράδειγμα, σε εξόρυξη δεδομένων για μια καμπάνια άμεσης αλληλογραφίας, η ομαδοποίηση μπορεί να αποκαλύψει ομάδες πελατών με ξεκάθαρα μοτίβα παραγγελιών. Γνωρίζοντας αυτά τα μοτίβα δημιουργούνται ευκαιρίες για προσωπικές αλληλογραφίες και διαφήμιση.

Τροποποίηση των δεδομένων δημιουργώντας, επιλέγοντας και μετατρέποντας μεταβλητές για να εστιάσετε την διαδικασία επιλογής μοντέλου.

Βασιζόμενοι στις ανακαλύψεις σας απο την φάση εξερεύνησης, μπορεί να χρειαστεί να χειριστείτε τα δεδομένα σας ώστε να περιλάβετε πληροφορίες όπως την ομαδοποίηση των πελατών και σημαντικών υποομάδων ή να εισάγετε νέες μεταβλητές. Μπορεί να χρειαστεί να κοιτάξετε για ακραίους και να μειώσετε τον αριθμό των μεταβλητών ώστε να περιοριστείτε στους πιο σημαντικούς. Μπορεί επίσης να χρειαστεί να τροποποιήσετε δεδομένα όταν εκείνα που εχουν εξορυχθεί μεταβάλλονται είτε λόγω νέων δεδομένων που γίνονται διαθέσιμα, είτε λόγω εσφαλμένων δεδομένων που έχουν πρόσφατα ανακαλυφθεί. Λόγω του ότι η εξόρυξη δεδομένων ειναι μια δυναμική και επαναλαμβανόμενη διαδικασία, μπορείτε να ενημερώσετε τις μεθόδους ή τα μοντέλα εξόρυξης δεδομένων όταν νέες πληροφορίες γίνονται διαθέσιμες.

Μοντελοποίηση των δεδομένων επιτρέποντας στο λογισμικό να αναζητήσει αυτόματα για έναν συνδυασμό δεδομένων που θα προβλέψει το επιθυμητό αποτέλεσμα με αξιοπιστία .

Οι τεχνικές μοντελοποίησης στην εξόρυξη δεδομένων περιλαμβάνουν τα νευρωνικά δίκτυα, δενδριτικά μοντέλα, λογιστικά μοντέλα και άλλα στατιστικά μοντέλα όπως η χρονολογικού τύπου ανάλυση, η αιτιολόγηση βασισμένη στην μνήμη και κύριες συνιστώσες. Κάθε τύπος μοντέλου έχει συγκεκριμένες δυνατότητες και είναι κατάλληλο για συγκεκριμένες συνθήκες εξόρυξης δεδομένων ανάλογα τα δεδομένα. Για παράδειγμα, τα νευρωνικά δίκτυα ειναι πολύ καλά στο να τοποθετούν πολύπλοκους μη γραμμικούς συσχετισμούς.

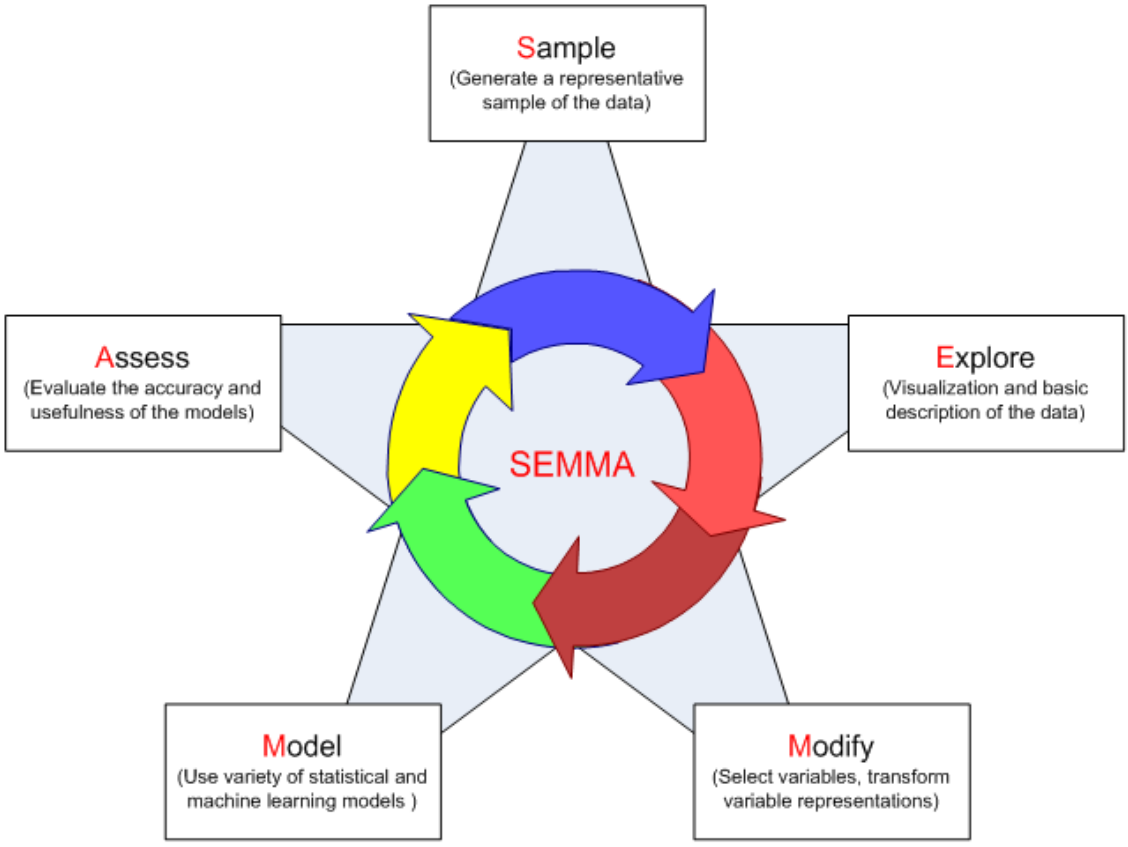
Αποκτήση πρόσβασης στα δεδομένα αποτιμώντας την χρησιμότητα και την αξιοπιστία των ευρυμάτων απο την διαδικασία εξόρυξης δεδομένων και εκτιμήστε πόσο καλά αποδίδει.

Ένας κοινός τρόπος εκτίμησης ενός μοντέλου είναι να το εφαρμόσετε ξεχωριστά σε ένα μέρος του συνόλου δεδομένων κατά την διαδικασία δειγματοληψίας. Έαν το μοντέλο είναι έγκυρο θα πρέπει να δουλεύει για αυτό το αποκλειστικό δείγμα αλλά και για το δείγμα που χρησιμοποιήθηκε για την δημιουργία του μοντέλου. Παρομοίως, μπορείτε να δοκιμάσετε το μοντέλο έναντι γνωστών δεδομένων. Για παράδειγμα, εάν γνωρίζετε ποιοι πελάτες του αρχείου είχαν υψηλά ποσοστά παρακράτησης και το μοντέλο σας προβλέπει την παρακράτηση, μπορείτε να ελέγξετε εάν το μοντέλο επιλέγει αυτούς τους πελάτες με ακρίβεια. Επιπροσθέτως, πρακτικές εφαρμογές αυτού του μοντέλου, όπως οι μερικές αλληλογραφίες σε μια καμπάνια άμεσης αλληλογραφίας, βοηθά στο να αποδείξει την εγκυρότητα του.

Με την αξιολόγηση των αποτελεσμάτων που αποκτήθηκαν απο κάθε στάδιο της διαδικασίας sEMMA, μπορείτε να καθορίσετε πως θα μοντελοποιήσετε νέες ερωτήσεις που

μπορεί να προκύψουν απο τα προηγούμενα αποτελέσματα και έτσι να μεταβείτε πίσω στην φάση εξερεύνησης για περαιτέρω βελτίωση των δεδομένων.

Μόλις εξελίξετε το μοντέλο ‘πρωταθλητή’ έχοντας χρησιμοποιήσει την προσέγγιση εξόρυξης sEMMA, θα πρέπει αυτό να αναπτυχθεί έτσι ώστε να σκοράρει νέες περιπτώσεις πελατών. Η ανάπτυξη του μοντέλου είναι το τελικό αποτέλεσμα της εξόρυξης δεδομένων, η τελική φάση κατά την οποία συνηδειτοποιείται το αποτέλεσμα της επένδυσης απο την διαδικασία εξόρυξης. Η επιχείρηση εξόρυξης αυτοματοποιεί την φάση εξέλιξης προμηθεύοντας κώδικα σκοραρίσματος σε γλώσσες προγραμματισμού SAS,C,Java και PMML. Δεν συλλαμβάνει τον κώδικα μόνο για αναλυτικά μοντέλα αλλά και για δραστηριότητες επεξεργασίας. Μπορείτε απρόσκοπτα να σκοράρετε τα δεδομένα παραγωγής σε ένα διαφορετικό μηχάνημα και να αναπτύξετε τον κώδικα σκοραρίσματος σε παρτίδες ή σε πραγματικό χρόνο στο διαδίκτυο ή άμεσα σε συσχετιζόμενες βάσεις δεδομένων. Αυτό έχει σαν αποτέλεσμα την ταχύτερη υλοποίηση και σας ελευθερώνει ώστε να ξοδέψετε περισσότερο χρόνο αποτιμώντας υπάρχοντα μοντέλα και αναπτύσσοντας τα (Dean, 2014)



Σχήμα 8: Αρχιτεκτονική της sEMMA (http://1.bp.blogspot.com/-h8jhlWeaLvs/T4y5UMwybjI/AAAAAAAAAHw/dKOzlU1Rm4o/s1600/data_mining_process_semma.png)

3.2.1 SEMMA ΓΙΑ ΤΗΝ ΕΠΟΧΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Πώς η μεθοδολογία sEMMA επέδρασε στην εποχή των μεγάλων δεδομένων; Η σύντομη απάντηση είναι πως όχι. Η sEMMA είναι μια λογική διαδικασία που μπορεί να ακολουθηθεί ανεξαρτήτως μεγέθους δεδομένων ή πολυπλοκότητας. Ωστόσο, το 's', ή δείγμα, στην λέξη sEMMA είναι λιγότερο καθοριστικό με τα σημερινά ισχυρά συστήματα που είναι διαθέσιμα για εξόρυξη δεδομένων.Πολύ μεγάλες και αναλυτικές βάσεις δεδομένων μπορούν να αντιμετωπιστούν με την χρήση sEMMA. (Dean, 2014)

3.3 ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Η δυαδική ταξινόμηση είναι το πιο κοινό είδος μοντέλου πρόβλεψης. Βασικοί φορείς λήψης αποφάσεων για επιχειρήσεις και άλλους οργανισμούς συχνά πρέπει να λάβουν κρίσιμες αποφάσεις γρήγορα, απαιτώντας από το σύστημα να πάρει μια θετική ή αρνητική απόφαση με σιγουριά. Αυτά τα συστήματα δεν είναι σχεδιασμένα να κάνουν πράγματα ελλιπώς ή χωρίς απόλυτη βεβαιότητα. Είτε τα κάνουν είτε όχι. Αυτό αποτυπώνεται πολύ καλά στην περίφημη λίστα ελέγχου 'πάω ή δεν πάω για έναρξη' όπου κάθε ελεγκτής πτήσεως ενός συστήματος πρέπει να απαντήσει. Η έναρξη μπορεί να προχωρήσει μόνο αφού ο διευθυντής της πτήσης έχει ανατρέξει σε όλα τα συστήματα και έχει λάβει την απάντηση 'πήγαινε για πτήση'. Στον κόσμο των επιχειρήσεων, πολλές αποφάσεις είναι δυαδικές, τέτοιες όπως για παράδειγμα αν θα σου επεκτείνω την πίστωση για αγορά αυτοκινήτου ή εάν θα ανταποκριθείς σε αυτή την εμπορική καμπάνια.

Με την προγνωστική μοντελοποίηση ενός δυαδικού στόχου, η πιθανότητα να συμβεί ή να μην συμβεί ένα γεγονός είναι μια επίσης πολύ χρήσιμη πληροφορία. Επιτρέψετε μου να σας δώσω ένα παράδειγμα για να σας διευκρινίσω αυτό το σημείο. Με έναν δυαδικό στόχο δημιουργούμε μια κατάσταση άσπρου ή μαύρου. Θα βρέξει αύριο – ναι ή όχι; Είτε θα βρέξει είτε δεν θα βρέξει, αλλά είναι άκρως απίθανο να ακούσεις τον μετεωρολόγο να δίνει πιθανότητες να βρέξει 0% ή 100% διότι ενώ μπορεί να βρέξει 0% ή 100%, η εκτίμηση είναι αναμφιβόλως ο βαθμός βεβαιότητας που έχουμε στην πρόβλεψή μας. Η εκτίμηση βεβαιότητας είναι συχνά πολύ πιο χρήσιμη από την ίδια την δυαδική πρόβλεψη. Μεταβάλλεται η συμπεριφορά σας εάν δείτε 10% πιθανότητες να βρέξει σε σχέση με το αν δείτε 40%; Προτίθεστε να αφήσετε τα παράθυρα ανοιχτά ή την οροφή του κάμπριο αυτοκινήτου σας κατεβασμένη; Και οι δύο προβλέψεις, λόγο του ότι το ποσοστό είναι λιγότερο του 50%, δείχνουν ότι είναι λιγότερο πιθανό να βρέξει από το να μην βρέξει. Παρ'όλα αυτά, με βρεγμένα καθίσματα στο αυτοκίνητό κατά την διάρκεια της διαδρομής προς το σπίτι σε μια ημέρα όπου ήταν απίθανο να βρέξει (30%), διατηρείς τα παράθυρά κλειστά εάν υπάρχει οποιαδήποτε περίπτωση να βρέξει διότι το πιθανό ρίσκο είναι πολύ μεγαλύτερο από το εν δυνάμει όφελος. Μια περίπτωση που βλέπουμε συχνά στην δυαδική προγνωστική μοντελοποίηση είναι όταν αντιμετωπίζουμε ένα σπάνιο γεγονός. Αυτό που

καθιστά ένα γεγονός σπάνιο διαφέρει αναλόγως την βιομηχανία και τον τομέα, αλλά είναι ευρέως αποδεκτό ότι το γεγονός αυτό συμβαίνει με εξαιρετικά χαμηλή πιθανότητα και γιαυτό τον λόγο μπορεί να απαιτεί ειδική μεταχείρηση στην ανάλυσή του.

Χρησιμοποιώντας ένα παράδειγμα από το σπορ του γκόλφ, ας εξετάσουμε το σπάνιο γεγονός με μεγαλύτερη προσοχή. Το πιο περιζήτητο συμβάν στο γκολφ είναι η τρύπα στο ένα. Είναι η στιγμή κατά την οποία ο γκολφέρ με ένα χτύπημα tee, κατά την διάρκεια τουλάχιστον ενός γύρου 9 τρυπών, πετυχαίνει την τρύπα με μια προσπάθεια. Εάν το καταφέρεις αυτό κατά την διάρκεια κάποιων τουρνουά, συχνά κερδίζεις σαν έπαθλο ενα αυτοκίνητο καθώς επίσης το έθιμο προστάζει πως ανεξαρτήτως του πότε συμβαίνει αυτό θα πρέπει να κεράσεις έναν γύρο ποτά σε όλα τα μέλη του κλάμπ. Για να είναι επίσημο (ή έγκυρο) η βολή πρέπει να γίνει παρουσία κάποιου άλλου. Ο σύνδεσμος γκόλφ των Ηνωμένων Πολιτειών κρατά αρχείο για όλες τις τρύπες του ενός και εκτιμά πως οι πιθανότητες να πετύχει κάποιος την τρύπα με την μια είναι 1 στις 33000. Επομένως, ενώ οι πιθανότητες για κάποιον να πετύχει μια τρύπα με την μία είναι πολύ μικρές (0.003%), ορισμένοι άνθρωποι, όπως ο Tiger Woods, οχι απλά πέτυχαν μια αλλά πολλές τρύπες με την μία. Ο Tiger έχει καταγεγραμμένες 18 τρύπες με την μια στο ενεργητικό του αλλά δεν ηγείται στην κατηγορία του. Οι περισσότερες τρύπες με την μια έχουν καταγραφεί από τον Norman Manley, 59 τον αριθμό. Για να καταλάβουμε το μέγεθος των πιθανοτήτων να πετύχει κάποιος τρύπα με την μία, θα μπορούσαμε να πούμε πως για κανέναν δεν είναι πιθανόν να πετύχει κάποια τρύπα με την μία. Υποθέτοντας πως 1 στα 33000 χτυπήματα καταλήγει σε μια τρύπα με την πρώτη και οτι υπάρχουν 14 τρύπες από τις οποίες θα μπορούσες να πετύχεις μια με την πρώτη, εάν έπαιζες έναν γύρο γκολφ κάθε ημέρα του χρόνου θα πετύχαινες κατά μέσο όρο μία τρύπα με την πρώτη κάθε επτά χρόνια. Άπαξ και έχουμε καθορίσει την πιθανότητα να συμβεί ένα γεγονός για κάθε παρατήρηση, τότε αυτά ταξινομούνται σε φθίνουσα σειρά από αυτά με τις μεγαλύτερες πιθανότητες σε εκείνα με τις μικρότερες, βασιζόμενα στην πιθανότητα να συμβεί το γεγονός. Χρησιμοποιώντας το άνωθεν παράδειγμα, εάν εμπορευόμασταν αναμνηστικές πλάκες τρυπών με την μια, θα εφαρμόζαμε το μοντέλο για όλους τους γκολφέρ στην βάση δεδομένων μας, ταξινομώντας την λίστα από τις περισσότερες έως τις λιγότερες πιθανότητες να πετύχουν κάποια τρύπα με την μία. Θα αποστέλναμε τότε μια επιστολή προσφοράς προς όλους τους γκολφέρ απο την κορυφή της λίστας και κάτω έως ότου ο προυπολογισμός της εμπορικής μας καμπάνιας εξαντληθεί ή προς εκείνους που υπερέβαιναν κάποιο όριο. Αυτό θα εξασφαλίσει πως θα στείλουμε στους γκολφέρ προσφορά για αναμνηστική πλάκα, βασιζόμενοι στο μοντέλο, ‘πιο πιθανόν να πετύχει κάποια τρύπα με την μία’ και οι οποίοι θα ενδιαφέρονταν για τις υπηρεσίες μας. (Dean, 2014)

3.4 ΠΟΛΥΕΠΙΠΕΔΗ ΤΑΞΙΝΟΜΗΣΗ

Η πολυεπίπεδη κατάταξη μοιάζει πολύ με την δυαδική κατάταξη με την εξαίρεση πως τώρα υπάρχουν περισσότερα από δύο επίπεδα. Η ονομαστική κατάταξη είναι μια επέκταση της δυαδικής κατάταξης. Υπάρχουν διάφορα παραδείγματα όπου η ονομαστική κατάταξη είναι κοινή αλλά για το μεγαλύτερο μέρος αυτός είναι ο πιο σπάνιος των στόχων. Παράδειγμα ενός τέτοιου μοντέλου μπορούμε να δούμε στην βιομηχανία κινητών τηλεφώνων όταν βλέπουμε τις απώλειες πελατών. Μια εταιρία μπορεί να μην ενδιαφέρεται μόνον για την δυαδική ανταπόκριση του κατά πόσον ένας λογαριασμός παραμένει ενεργός ή όχι. Αντίθετα, μπορεί να θέλει να ‘βουτήξει΄ βαθύτερα και να δει την ονομαστική ανταπόκριση εθελούσιας απώλειας (ο πελάτης επιλέγει να ακυρώσει το συμβόλαιο), μη εθελούσιας απώλειας (τερματισμός συμβολαίου λόγω παράλειψης πληρωμών), ή ενός ενεργού πελάτη.

Σε πολλές περιπτώσεις, προκύπτει πρόβλημα ονομαστικής κατάταξης όταν μια κατ’ εξαίρεση περίπτωση προστίθεται σε κάτι που θα μπορούσε να είναι μια δυαδική απόφαση. Αυτό συμβαίνει στην περίπτωση όπου για παράδειγμα εμποδίζεται κάποια απάτη μέσω πιστωτικής κάρτας. Όταν ξεκινά μια συναλλαγή με πιστωτική κάρτα, η εκδοτική αρχή της κάρτας έχει στην διάθεσή της ένα πολύ μικρό περιθώριο χρόνου για να αποδεχτεί ή να απορρίψει την συναλλαγή. Ενώ αυτό θα μπορούσε να θεωρηθεί ως ένα απλό δυαδικό πρόβλημα, αποδοχής ή απόρριψης, υπάρχουν κάποιες συναλλαγές όπου η απόφαση δεν είναι τόσο ξεκάθαρη και μπορεί να εμπίπτει εντός ορισμένων γκρίζων ζωνών. Ένα τρίτο επίπεδο ανταπόκρισης μπορεί να προστεθεί για να υποδηλώσει ότι η συγκεκριμένη συναλλαγή απαιτεί περαιτέρω εξέταση προτού υπάρξει απόφαση αποδοχής ή απόρριψης.

Στο πρόβλημα ονομαστικής κατάταξης ανακύπτουν κάποιες επιπρόσθετες επιπλοκές τόσο από υπολογιστικής όσο και από αναφορικής όψεως. Αντί απλά να υπολογίζεται η πιθανότητα ενός γεγονότος (P) και τότε να λαμβάνεται η 1-P ώστε να καταλήγουμε στην πιθανότητα ενός μη-γεγονότος, χρειάζεται να υπολογίσετε την πιθανότητα του γεγονότος #1 (P1), την πιθανότητα του γεγονότος #2 (P2) και ούτω καθεξής έως ότου το τελευταίο επίπεδο το οποίο μπορεί να υπολογιστεί χρησιμοποιώντας

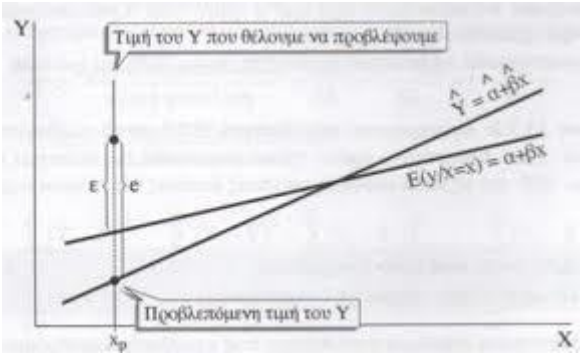
$$1 - \sum_{i=1}^{n-1} P$$

Υπάρχει επίσης μια πρόκληση στον υπολογισμό του ποσοστού εσφαλμένης ταξινόμησης. Αφού υπάρχουν πολλές επιλογές, η τιμές αναφοράς πρέπει να είναι ρυθμισμένες έτσι ώστε να μεταφράζονται εύκολα από τον αναγνώστη της αναφοράς. (Dean, 2014)

3.5 ΠΡΟΒΛΕΨΗ ΔΙΑΣΤΗΜΑΤΟΣ

Το τελευταίο είδος πρόβλεψης είναι η πρόβλεψη διαστήματος η οποία χρησιμοποιείται όταν το επίπεδο του στόχου είναι συνεχές στην αριθμητική γραμμή. Ο μισθός είναι ένα

παράδειγμα πρόβλεψης όπου τόσο οι εργοδότες όσο και οι υπάλληλοι θα ήθελαν να κάνουν με ακρίβεια. Η βιομηχανία ασφάλειας κατοικιών και ατυχημάτων είναι ένας τομέας με πολλά μοντέλα προβλέψεως διαστημάτων. Εάν είσαι κάτοχος αυτοκινήτου στις Ηνωμένες Πολιτείες, απαιτείται να το έχεις ασφαλισμένο. Για να αποκτήσεις την ασφάλεια, πιθανόν να ζήτησες προσφορές από διάφορες ασφαλιστικές εταιρίες και κάθε μία σου έδωσε ελαφρώς διαφορετική τιμή. Αυτή η διαφοροποίηση των τιμών οφείλεται στα διαφορετικά προγνωστικά μοντέλα και τους επιχειρηματικούς παράγοντες που χρησιμοποιεί η κάθε ασφαλιστική εταιρία. Οι επιχειρηματικοί παράγοντες είναι το μέγεθος έκθεσης σε ορισμένες γεωγραφικές αγορές ή μια συνολική οδηγία υπο την προσπάθεια απόκτησης μεριδίου σε συγκεκριμένη αγορά είτε γεωγραφικά είτε οικονομικά. Οι εταιρίες στην ασφαλιστική βιομηχανία, γενικώς χρησιμοποιούν τρείς διαφορετικούς τύπους μοντέλων πρόβλεψης διαστημάτων συμπεριλαμβανομένης της συχνότητας απαιτήσεων, της σοβαρότητας της κατάστασης και του καθαρού ασφαλίστρου. Η ασφαλιστική εταιρία θα κάνει προβλέψεις για κάθε ένα από αυτά τα μοντέλα βασιζόμενη στα ιστορικά τους δεδομένα και τις ιδιαίτερες πληροφορίες σου όπως την μάρκα και το μοντέλο του αυτοκινήτου σου, τα ετήσια χιλιόμετρα που διανύεις, το ιστορικό παραβάσεων στην οδήγησή σου και ούτω καθεξής. (Dean, 2014)



Σχήμα 9 :Πρόβλεψη διαστήματος(https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcR5H0o72Np0rOLU8CjVfX5XJIxptClgIrmj4XSq8Byi3u9YALlOzw)

3.6 ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ ΠΡΟΒΛΕΨΗΣ

Ένα από αυτά που θα πρέπει να ληφθούν υπ'όψιν κατά την διαδικασία δημιουργίας ενός μοντέλου πρόβλεψης είναι να καθορίσετε πιο μοντέλο είναι καταλληλότερο. Ένα μοντέλο αποτελείται από όλους τους μετασχηματισμούς , τις κατηγορίες, τις επιλογές μεταβλητών και όλους εκείνους τους χειρισμούς που εφαρμόζονται στα δεδομένα επιπρόσθετα του επιλεγμένου αλγορίθμου και των σχετικών του παραμέτρων. Ο μεγάλος αριθμός επιλογών και συνδυασμών καθιστά την μέθοδο εξαναγκασμού 'ανεύρεσης των πάντων' πρακτικά αδύνατη για οποιοδήποτε πρόβλημα εξόρυξης δεδομένων. Επομένως το ζήτημα της αξιολόγησης δεδομένων παίζει έναν σημαντικό ρόλο για την επιλογή του καλύτερου

μοντέλου. Με απλά λόγια, η αξιολόγηση του μοντέλου είναι η προσπάθεια ανεύρεσης του καλύτερου μοντέλου για την εφαρμογή σας σύμφωνα με τα υπάρχοντα δεδομένα. Η πολυπλοκότητα έγκειται στον όρο ‘καλύτερο΄. Όπως όταν αγοράζουμε Ένα πλυντήριο, όπου πρέπει να το εξετάσουμε από διάφορες απόψεις και δεν μπορούν όλοι να συμφωνήσουν στο τι είναι το ‘καλύτερο’. Η κοινή δέσμη μέτρων αξιολόγησης ενός μοντέλου απαριθμούνται και ορίζονται παρακάτω. Για παράδειγμα, θεωρήστε μια τόπική φιλανθρωπική οργάνωση που διοργανώνει μια δράση τροφίμων. Διαθέτει μια λίστα δωρητών 125000 ατόμων και δεδομένα από προηγούμενες εκστρατείες που θα χρησιμοποιείσει για να εκπαιδεύσει το μοντέλο. Διαχωρίζουμε τα δεδομένα έτσι ώστε 100.000 άνθρωποι βρίσκονται στην ομάδα εκπαίδευσης και 25.000 βρίσκονται στην ομάδα επικύρωσης. Και οι δυο ομάδες έχουν ποσοστό ανταπόκρισης 10%. Η ομάδα επικύρωσης θα χρησιμοποιηθεί για αξιολόγηση, η οποία είναι η βέλτιστη πρακτική. (Dean, 2014)

Table 4.1 Decision Matrix for Model Assessment

	Predicted Nonevent	Predicted Event
Nonevent	True negative	False positive
Event	False negative	True positive

3.7 ΤΑΞΙΝΟΜΗΣΗ

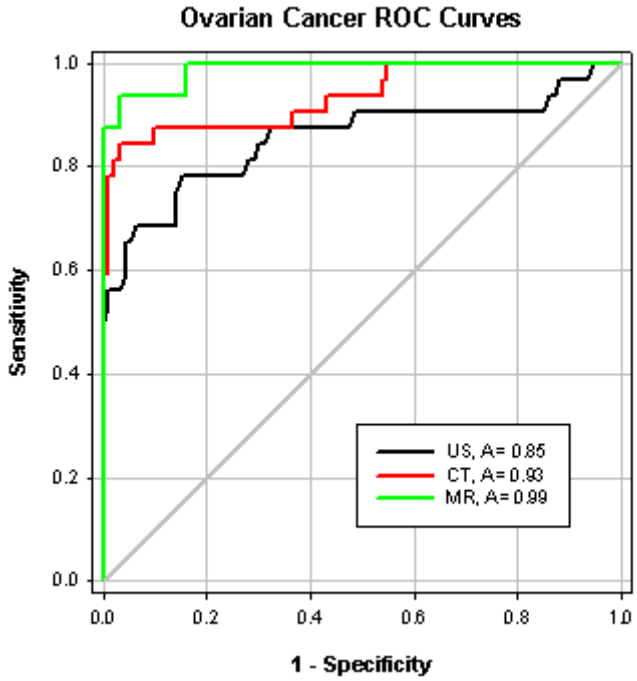
Η ταξινόμηση είναι μια δημοφιλής μέθοδος. Με αυτή είναι εύκολο να εξηγηθεί αυτό που οι περισσότεροι άνθρωποι θεωρούν ως ‘καλύτερο’ μοντέλο και μετρά την ικανότητα του μοντέλου στο εύρος όλων των τιμών. Εάν η αναλογία των γεγονότων και μη-γεγονότων δεν είναι περίπου ίση, τότε οι τιμές πρέπει να προσαρμοστούν έτσι ώστε να λαμβάνονται ορθές αποφάσεις. (Dean, 2014)

3.8 ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΛΕΙΤΟΥΡΓΙΑΣ ΔΕΚΤΗ

Τα χαρακτηριστικά της λειτουργίας του δέκτη (ROC) υπολογίζονται για όλα τα σημεία και εμφανίζονται γραφικά προς ερμηνεία. Οι άξονες του διαγράμματος ROC είναι η ‘Ευαισθησία’ και η ‘1-Ιδιαιτερότητα’ που υπολογίστηκαν από τα ποσοστά ταξινόμησης.

Table 4.2 Formulas to Calculate Different Classification Measures

Measure	Formula
Classification rate (accuracy)	$\frac{true\ negative + true\ positive}{total\ observations} \times 100$
Misclassification rate	$(1 - \frac{true\ negative + true\ positive}{total\ observations}) \times 100$
Sensitivity (true positive rate)	$\frac{true\ positive}{true\ positive + false\ negative} \times 100$
Specificity (true negative rate)	$\frac{true\ negative}{false\ positive + true\ negative} \times 100$
1-Specificity (false positive rate)	$\frac{false\ positive}{false\ positive + true\ negative} \times 100$



Σχήμα 10 :Αναπαράσταση ROC γραφήματος
(http://www.sigmaplot.com/products/sigmaplot/productuses/roc_images/image013.gif)

3.9 ΑΝΥΨΩΣΗ

Ανύψωση (lift) είναι η αναλογία του ποσοστού των ορθά αποκρινόμενων με το ποσοστό της αρχικής απάντησης. Για να υπολογίσουμε την ανύψωση θα πρέπει να συνοδεύεται από μια ποσόστωση στα δεδομένα. Αυτό συχνά αναφέρεται ως βάθος αρχείου και συνήθως επιλέγεται το πρώτο ή το δεύτερο δέκαδικό. Για το παράδειγμα της δράσης τροφίμων, εάν

υπολογίσουμε την ανύψωση στο πρώτο δεκαδικό (10% των δεδομένων), το αρχικό (ή το τυχαίο) μοντέλο θα πρέπει να έχει 2500 αποκρινόμενους στην εκστρατεία έτσι ώστε στο πρώτο δεκαδικό να υπάρχουν 250 αποκρινόμενοι (2500*0,1). Το μοντέλο μας είναι καλό. Συλλαμβάνει 300 αποκρινόμενους στο πρώτο δεκαδικό οπότε η ανύψωση στο πρώτο δεκαδικό ειναι 1.2 (300/2500=12% ληφθήσα απάντηση /10% αρχική απάντηση). Προτιμούμε να χρησιμοποιούμε αθροιστική ανύψωση για την αξιολόγηση του μοντέλου μας διότι είναι μονοτονική και στην πράξη οι εκστρατείες ταξινομούν μια λίστα σύμφωνα με την πιθανότητα ανταπόκρισης και μόνο τότε θα έρθει στην αγορά, μέχρις ότου παρατηρηθεί μια φυσική διάλυση ή εξαντληθεί ο πρυπολογισμός της εκστρατείας. (Dean, 2014)

3.10 ΚΕΡΔΟΣ

Το κέρδος μοιάζει πολύ με την ανύψωση με την διαφορά οτι το 1 αφαιρείται απο την τιμή % για το δεδομένο δεκαδικό. Για το παράδειγμα της δράσης τροφίμων το κέρδος θα είναι 0.2 στο πρώτο δεκαδικό.

3.11 ΤΟ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ ΤΟΥ AKAIKE

Το κριτήριο πληροφορίας του Akaike (AIC) είναι ένα στατιστικό μέτρο της καταλληλότητας ενός συγκεκριμένου μοντέλου. Προκειμένου να καθορίσουμε την καλύτερη εξειδίκευση του μοντέλου μας και να προσδιορίσουμε την άριστη χρονική υστέρηση, χρησιμοποιούμε το κριτήριο πληροφοριών Akaike (Akaike Information Criteria). Το κριτήριο αυτό ορίζεται ως εξής:

Μεγιστοποιεί την έκφραση $-2(LL + k)$ όπου

k = αριθμός εκτιμόμενων παραμέτρων (για γραμμική παλινδρόμηση ο αριθμός των όρων σε λειτουργία)

LL= μέγιστη τιμή της συνάρτησης λογαριθμικής πιθανότητας για το δεδομένο μοντέλο.

Όσο μικρότερο το μοντέλο, τόσο καλύτερα το μοντέλο προσαρμόζει τα δεδομένα. Εξαιτίας του όρου ‘k’ ο μικρότερος αριθμός των παραμέτρων του μοντέλου ευνοείται. Οι τιμές του AIC μπορούν να είναι αρνητικές αλλά δεν θυμάμαι ποτέ να αντιμετώπισα στην πράξη αυτό το το πρόβλημα, στην εξόρυξη δεδομένων. (Dean, 2014)

3.12 ΤΟ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ BAYESIAN

Το κριτήριο πληροφορίας Bayesian (BIC) είναι ένα στατιστικό παρόμοιο με το AIC με την διαφορά Ότι μεγιστοποιεί την έκφραση $-2LL + k \times ln(n)$ όπου:

n = αριθμός παρατηρήσεων (ή μεγεθος δείγματος)

k = αριθμός εκτιμόμενων παραμέτρων (για γραμμική παλινδρόμηση ο αριθμός των όρων σε λειτουργία)

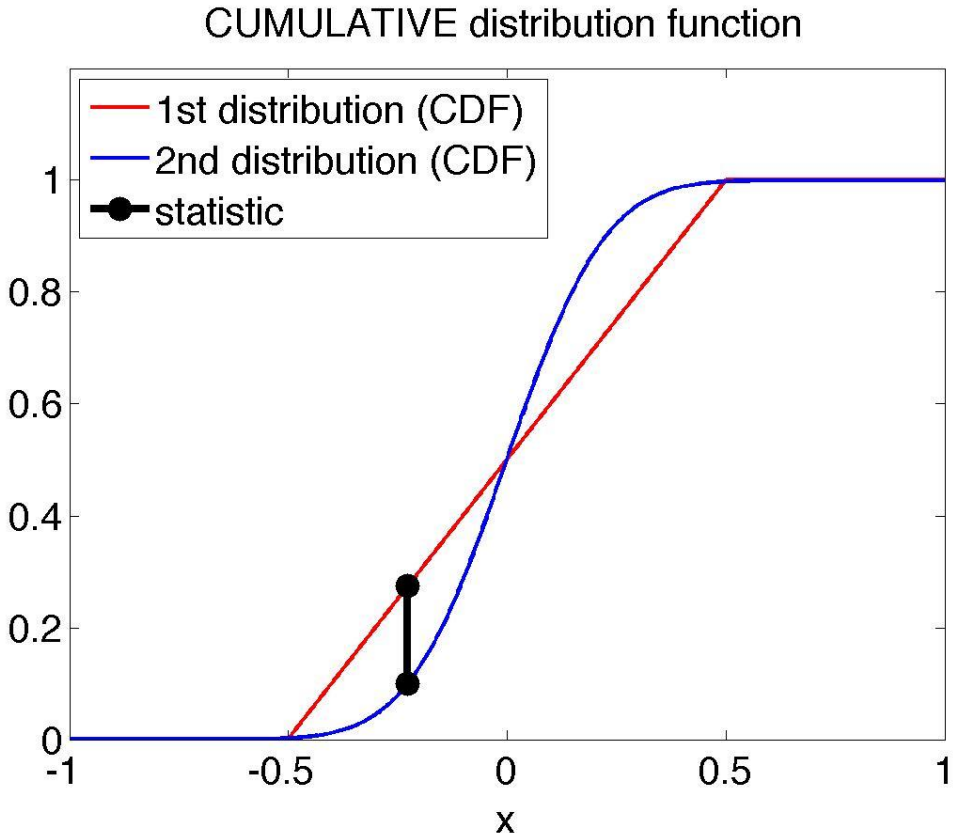
LL = μέγιστη τιμή της συνάρτησης λογαριθμικής πιθανότητας για το δεδομένο μοντέλο.

Για οποιόδηποτε από τα δυο μοντέλα, εκείνο με τον χαμηλότερο BIC είναι το καλύτερο. Αυτό συμβαίνει διότι ο αριθμός των όρων στο μοντέλο είναι μικρότερος, οι μεταβλητές του μοντέλου ερμηνεύουν καλύτερα την διαφοροποίηση του στόχου, ή συμβαίνουν και τα δύο. (Dean, 2014)

3.13 KOLMOGOROV-SMIRNOV

Η στατιστική Kolmogorov-Smirnov (KS) δείχνει το σημείο του μέγιστου διαχωρισμού της ευαισθησίας του μοντέλου και το σημείο εκκίνησης της καμπύλης ROC.

Η αξιολόγηση του μοντέλου, ιδανικά γίνεται σε ένα διαχωρισμένο μέρος που είναι αντιπροσωπευτικό των δεδομένων αλλά δεν χρησιμοποιήθηκε στην φάση δημιουργίας του μοντέλου. Αυτό το μέρος (συχνά αποκαλείται χώρισμα επικύρωσης ή δοκιμής) είναι ουσιώδες για να μετρήσετε πόσο καλά γενικεύει το μοντέλο σας σε νέα εισερχόμενα δεδομένα. (Dean, 2014)



Σχήμα 11:Συνάρτηση Αθροιστικής Κατανομής (http://ivrl.epfl.ch/files/content/sites/ivrg/files/supplementary_material/AL_ACMMM12/webpage_supplementary_material/statistic/KS2.jpg)

3.14 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η Ανάλυση παλινδρόμησης είναι πιθανόν η πρώτη μέθοδος προγνωστικής μοντελοποίησης που μαθαίνεις ως επαγγελματίας κατά τη διάρκεια των σπουδών σου. Η έννοια της

Παλινδρόμησης δημοσιεύθηκε για πρώτη φορά στις αρχές του 1800 από τον Adrien-Marie Legrendre και Carl Gauss.O Legrendre γεννήθηκε σε μια πλούσια γαλλική οικογένεια και συνέβαλε σε μια σειρά από εξελίξεις στους τομείς των μαθηματικών και της στατιστικής.O Gauss, σε αντίθεση, γεννήθηκε σε μια φτωχή οικογένεια στη Γερμανία. Ήταν ένα παιδί-θαύμα στα μαθηματικά, αλλά σε όλη του τη ζωή ήταν απρόθυμος να δημοσιεύει οποιαδήποτε εργασία του επειδή ένιωθε ότι δεν ήταν υπεράνω κριτικής. Μετά το θάνατό του,σε πολλά από τα προσωπικά του ημερολόγια ανακαλύφθηκαν λεπτομέρειες των ιδεών του και των σκέψεων του. Οι ιδέες αυτές ήταν σημαντικές πρόοδοι στον τομέα των μαθηματικών. Εκτιμάται ότι εάν Gauss ήταν πιο επιθετικός στην δημοσιεύσει της δουλειά του, θα μπορούσε να προχωρήσει στον τομέα των μαθηματικών περισσότερα από 50 χρόνια. Οι δύο αυτοί λαμπροί άντρες μαζί με τον Sir K.A. Fischer στις αρχές του εικοστού αιώνα περιγράφουν τη μέθοδο της παλινδρόμησης. Ο όρος «παλινδρόμηση» επινοήθηκε από τον Francis Galton (ο οποίος ήταν εξάδελφος του Κάρολου Δαρβίνου) για να περιγράψει τη βιολογική διαδικασία της ακραίας τιμής. (Dean, 2014)

Η παλινδρόμηση είναι μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών . Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, αλλιώς καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, είναι να αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψη τις τιμές της κατηγορίας για τα νέα δεδομένα. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο προϊον ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος. Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (όπως π.χ. βάρος, ταχύτητα ή ηλικία). Αντίθετα, η παλινδρόμηση δεν λειτουργεί καλά με κατηγορικά δεδομένα (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

<u>Βασικό Παράδειγμα Τακτικής Ελαχίστων Τετραγώνων</u>

Ως παράδειγμα, ας εξετάσουμε τη σχέση του βάρους και ύψους για μια τάξη 19 μαθητών γυμνασίου. Υπάρχει μια αναμενόμενη και φυσική σχέση, ένας θετικός συσχετισμός μεταξύ του βάρους και του ύψους όπου αναμένουμε από τους μαθητές που ζυγίζουν περισσότερο να είναι ψηλότερη. Ας δείξουμε πρώτα μια πλοκή με το βάρος και το ύψος για να δούμε τι σχέση μοιάζει. Όπως μπορείτε να δείτε στο Σχήμα 12 υπάρχει μια τάση ότι όσο πιο πολύ

ζυγίζεις ,τόσο πιο ψηλός είσαι αλλά δεν υπάρχει σε καμία περίπτωση τέλεια σχέση. Θέλουμε να βρούμε μια γραμμή που προβλέπει καλύτερα το ύψος του φοιτητή γνωρίζοντας το βάρος του / της.

όπως στην επόμενη εξίσωση:

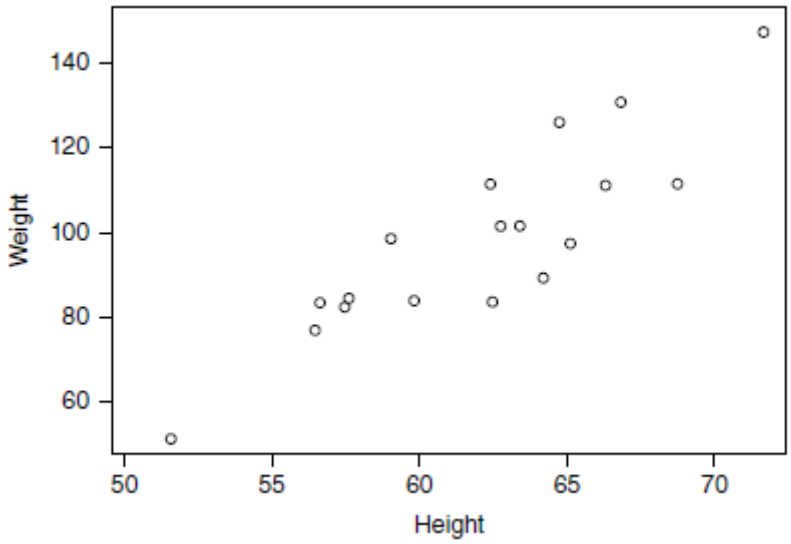
$$Hi = \beta 0 + wi * \beta i$$

Όπου :

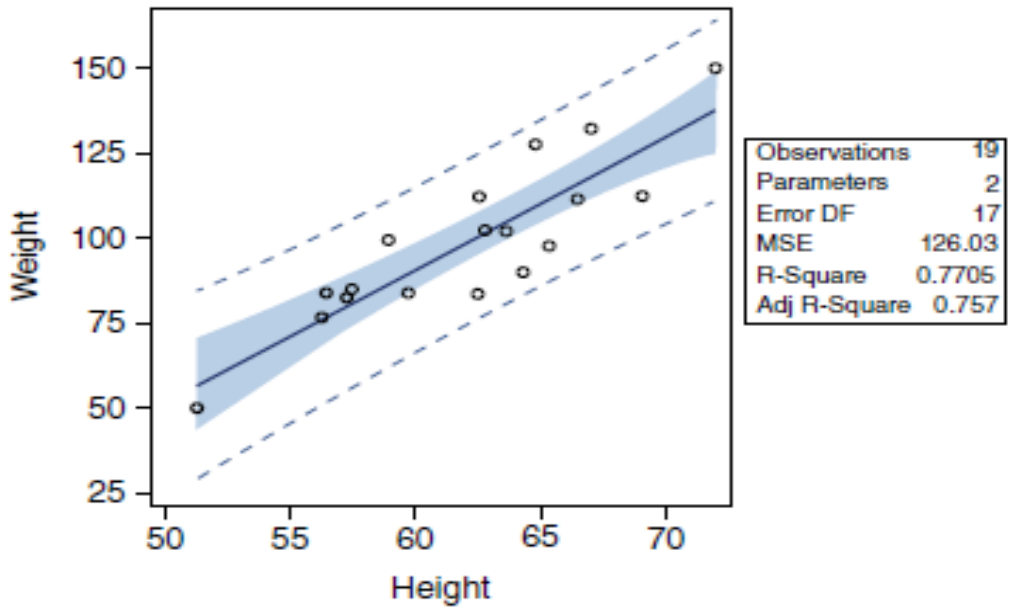
Hi = ύψος ενός μαθητή

Wi = βάρος ενός μαθητή

$\beta 0$ = σημείο τομής



Σχήμα 12 :Διάγραμμα διασποράς του ύψους και του βάρους για την κατηγορία



Σχήμα 13:Κατάλληλη διασπορά για το βάρος

Στην ανάλυση παλινδρόμησης, η καλύτερη γραμμή είναι εκείνη που ελαχιστοποιεί το άθροισμα των σφαλμάτων σε όλα τα σημεία δεδομένων. Το σφάλμα για μια παρατήρηση μετράται ως η τετραγωνική απόσταση μεταξύ της προτεινόμενης γραμμής, που είναι η προβλεπόμενη τιμή, και ο ανοιχτός κύκλος, η οποία είναι η πραγματική της αξίας. Γιατί είναι η κατακόρυφη απόσταση μεταξύ του σημείου και η τετράγωνο γραμμή; Επειδή με τα σημεία πάνω από τη γραμμή και άλλα σημεία κάτω από τη γραμμή, μια απλή προσθήκη της απόστασης μεταξύ των σημείων θα οδηγήσει σε έναν άπειρο αριθμό των γραμμών που θα μας δώσει μια αθροίζουσα τιμή 0. Αλλά χρησιμοποιώντας το τετράγωνο της απόστασης υπάρχει μία ενιαία καλύτερη γραμμή που μας δίνει το ελάχιστο άθροισμα των τετραγώνων. Αυτή η στατιστική είναι που ονομάζεται το άθροισμα των τετραγώνων, και αυτό είναι ένα πολύ χρήσιμο μέτρο του πόσο καλά η γραμμή παλινδρόμησης ταιριάζει με τα δεδομένα. Μπορείτε να δείτε τις καλύτερες γραμμές παλινδρόμησης.

Εκτός από τον υπολογισμό του αθροίσματος των τετραγώνων για να βρείτε την καλύτερη τοποθέτηση γραμμής, ένας αριθμός των διαγνωστικών πρέπει να θεωρηθεί ως μέρος της εκτέλεση μιας ανάλυσης παλινδρόμησης. (Dean, 2014)

3.14.1ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η Λογιστική Παλινδρόμηση είναι μία τεχνική για την πραγματοποίηση ανάλυσης δεδομένων που αφορούν την μελέτη και την πρόβλεψη τιμών κάποιας κατηγορικής εξαρτημένης μεταβλητής και χρησιμοποιεί ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές .Συνήθως τέτοιες αναλύσεις συναντάμε στους χώρους της υγείας για την μελέτη της

θεραπείας ή όχι των ασθενών, του Marketing για την αγορά ή όχι κάποιων προϊόντων, της παιδείας για την επιτυχία ή όχι των μαθητών στις εξετάσεις. Για δύο βασικούς λόγους δεν μπορεί να πραγματοποιηθεί μέσω του αλγορίθμου της Γραμμικής Παλινδρόμησης η μελέτη της σχέσης της κατηγορικής εξαρτημένης μεταβλητής.

Πρώτον, όταν προβλέπουμε τις τιμές μία κατηγορικής εξαρτημένης μεταβλητής, στην ουσία υπολογίζουμε την πιθανότητα με την οποία η εξαρτημένη μεταβλητή θα λάβει κάποια συγκεκριμένη τιμή. Η τιμή της πιθανότητας αυτής , εξ ορισμού,θα πρέπει να παίρνει τιμές μεταξύ του 0 και του 1. Με την χρήση όμως της Γραμμικής Πολλαπλής Παλινδρόμησης μπορούμε να υπολογίσουμε τιμές πιθανότητας μεγαλύτερες του 1 ή μικρότερες του 0, δηλαδή άτοπο .

Δεύτερον, η πολλαπλή γραμμική παλινδρόμηση θα πρέπει να ικανοποιεί την υπόθεση της ισότητας των διακυμάνσεων.Στην περίπτωση που η εξαρτημένη μεταβλητή είναι διχ οτομική, έχει τυπική απόκλιση $(St\ dev)\ (p)(\sqrt{1-p})$, όπου ρ είναι η μέση τιμή της μεταβλητής.Η ομοιογένεια της διακύμανσης των τιμών της εξαρτημένης μεταβλητής δεν είναι δυνατόν να ικανοποιείται λόγω της συναρτησιακής σχέσης της τυπικής απόκλισης με την μέση τιμή,

<u>Η ΕΞΙΣΩΣΗ ΤΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.</u>

Η πιο διαδεδομένη, έκφραση της εξίσωσης της Λογιστικής Παλινδρόμησης είναι: $ln(odds) = a + b1\,x\,1 + b2\,x2 \ldots . + bk\,xk$.Το δεξί μέρος της δημιουργείται από ένα γραμμικό συνδυασμό ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο της παλινδρόμησης. Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με την μορφή του λογαρίθμου των odd: $odds = prob/(1-prob)$. Εναλλακτικά το odds ονομάζεται και logt ενώ ο όρος Prob εκφράζει την πιθανότητα να συμβεί το γεγονός που έχει ορισθεί σαν επιτυχία του πειράματος. Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση της παλινδρόμησης εκτιμούνται βάση της μεθόδου Μέγιστης Πιθανοφάνειας βάση της μεθόδου αυτής η τιμή των συντελεστών των ανεξάρτητων μεταβλητών είναι αυτή που κάνει τις παρατηρηθήσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάση του σετ των ανεξάρτητων μεταβλητών.

Βήματα δημιουργίας του μοντέλου της Λογ. Παλινδρ/σης

Τα βήματα κατασκευής του μοντέλου της Λογιστικής Παλινδρόμησης είναι ανάλογα αυτών της γραμμικής παλινδρόμησης.

- Προσδιορίζουμε το μέγεθος του ενδιαφέροντος (εξαρτημένη μεταβλητή) και το σετ των ανεξάρτητων μεταβλητών που θα συμμετέχουν στην παλινδρόμηση.
- Διερευνούμε τα δεδομένα για τυχόν ύπαρξη ασυνήθιστων κινήσεων όπως, ακραίες τιμές, ελλείπουσες τιμές κ. λ. π.

- Ελέγχουμε την ικανοποίηση των υποθέσεων για την σωστή εφαρμογή της Λογιστικής Παλινδρόμησης.
- Δημιουργούμε την εξίσωσης της παλινδρόμησης.
- Μελετάμε την επίδραση κάθε ανεξάρτητης μεταβλητής στο μοντέλο.
- Εξετάζουμε την ικανοποίηση των υποθέσεων της Τεχνικής και διερευνούμε την πιθανότητα κάποια συγκεκριμένη τιμή να επηρεάζει υπερβολικά τα αποτελέσματα. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι για την σωστή εφαρμογή της Λογιστικής Παλινδρόμησης απαιτείται μεγάλο δείγμα, προκειμένου να έχουμε ένα αξιόπιστο αποτέλεσμα. Ένας εμπειρικός κανόνας αναφέρει ότι το δείγμα θα πρέπει να είναι 30 φορές μεγαλύτερο από το αριθμό των παραμέτρων που εκτιμά το μοντέλο. Επιπλέον, σε περίπτωση που ενδιαφερόμαστε να χρησιμοποιήσουμε το μοντέλο για πρόβλεψη θα πρέπει να αξιολογήσουμε την αποτελεσματικότητά του. Αυτό σημαίνει ότι δημιουργούμε την εξίσωση σε ένα μέρος των δεδομένων και σε ένα επόμενο βήμα ελέγξουμε την αποτελεσματικότητά της, στο υπόλοιπο δείγμα

3.15 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Εισαγωγή

Τα Τεχνητά Νευρωνικά Δίκτυα είναι μια μορφή τεχνητής νοημοσύνης τα οποία αποτελούν μια προσπάθεια προσέγγισης της λειτουργίας του ανθρώπινου εγκεφάλου. Είναι απλοποιημένα μοντέλα του κεντρικού νευρικού συστήματος του ανθρώπου. Μέχρι σήμερα έχουν προσαρμοστεί επιτυχημένα σε ένα ευρύ φάσμα περιοχών για την επίλυση προβλημάτων ταξινόμησης ή πρόβλεψης ,όπως η βιολογία ,η ιατρική ,η γεωλογία .Αποτελούνται από διασυνδεδεμένα υπολογιστικά στοιχεία που έχουν την ικανότητα να ανταποκρίνονται σε ερεθίσματα που δέχονται την είσοδο τους και να μαθαίνουν να προσαρμόζονται στο περιβάλλον τους .

Η έρευνα για τα Τεχνήτα Νευρωνικά Δίκτυα είναι εμπνευσμένη από την λειτουργία και την δόμη του εγκεφάλου. Οι νευρώνες είναι το δομικό στοιχείο του εγκεφάλου και του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελείται ένας υπολογισμός με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρείς τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων.

Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

Τα ΤΝΔ χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων ώστε να διατηρούν τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία . Τα ΤΝΔ έχουν ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ακόμα και τα απλά μοντέλα μπορούν να δημιουργήσουν ενδιαφέροντα δίκτυα αρκεί να έχουμε δυο βασικά χαρακτηριστικά:

1. Πλαστικότητα των νευρώνων δηλαδή οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους έτσι ώστε να διευκολύνεται η διαδικασία της μάθησης
2. Το δίκτυο πρέπει να αποτελείται από πολλούς νεωρώνες έτσι ώστε να έχουμε παραλληλισμό της επεξεργασίας και κατανομή της πληροφορίας.

Τα ΤΝΔ μοιάζουν με τον εγκέφαλο στα εξής σημεία :

1. Η γνώση αποκτάται από το δίκτυο μέσα από μια διαδικασία μάθησης-εκπαίδευσης
2. Η γνώση αποθηκεύεται στις δυνάμεις σύνδεσης των νευρώνων,οι οποίες είναι τα συναπτικά βάρη. (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

3.15.1 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

Τα Νευρωνικά Δίκτυα είναι μια νέα περιοχή σχετικά και ουσιαστικά δεν υπάρχει μεγάλη προιστορία, όπως σε άλλες επιστήμες .Σε διεθνές επίπεδο ξεκίνησε κατά τις τελευταίες δεκαετίες ,αλλά το 1980 δόθηκε η μέγαλη ώθηση .Σ΄ αυτό βοήθησε η μεγάλη ανάπτυξη του υλικού των Η/Υ αλλά και η ανάπτυξη αλγορίθμων εκπαίδευσης.

Το 1943 παρουσιάσθηκε το πρώτο μοντέλο Νευρωνικού Δικτύου από τους McCulloch και Pitts το οποίο πρότεινε ότι οι νεωρώνες είναι η βασική μόναδα του δικτύου.Στην εργασία τους παρουσίασαν για πρώτη φορά την ιδέα ότι ένα Νευρωνικό Δίκτυο αποτελείται από ένα σύνολο ενός μέγαλου αριθμού νευρωνών και έδειξαν πώς θα μπορούσαν να λειτούργουν οι νευρωνές με τις διασυνδέσεις τους.Αυτή θεωρείται η πρώτη εικόνα ενός Νερωνικού δικτύου.

Το 1974 ,οι ίδιοι συγγραφείς προχώρησαν σε ένα πιο εξελιγμένο πρότυπο για την αναγνώριση σχημάτων.Σ αυτό το πρότυπο ,ο νεωρώνας θεωρείται ότι μπορεί να έχει μόνο δύο καταστάσεις ,οί οποίες είναι ότι είτε πυροδοτεί ή βρίσκεται σε ηρεμία. Μπορεί να έχει πολλές εισόδους αλλά δίνει μία μόνο έξοδο .Οι έξοδοι από διαφορετικούς νευρώνες δεν επιτρέπει να ενώνονται αλλά πρεπει υποχρεωτικά να οδηγούν σε είσοδο άλλου νεωρώνα.Οι απολήξεις των νευρώνων είναι δύο ειδών :διεργετικές και ανασταλτικές. Όταν ο νευρώνας πυροδοτεί ,στέλνει ένα παλμό.Οι λειτουργίες αυτές γίνονται πάντα σε διακριτό χρόνο και υποτίθεται ότι όλοι οι νεωρωνες αποκρίνονται ταυτόχρονα ,δηλαδή το σύστημα δρα συγχρονισμένα.Τ α δίκτυα McCulloch-Pitts προσπαθούν να εξηγήσουν για πρώτη φορά πως δουλεύει η μνήμη. Θεωρούν ότι ένας πιθανός μηχανισμός μνήμης μπορεί να είναι η ύπαρξη

κλειστών διαδρομών του σήματος μέσα στο δίκτυο.Ετσι ,μια ίνα ενώνει την έξοδο ενός κυττάρου με το σημείο εισόδου στο ίδιο κύτταρο ,δημιουργώντας έναν μηχανισμό ανάδρασης (feedback)

Τις εργασίες των McCulloch-Pitts χρησιμοποίησε λίγα χρόνια αργότερα Ο J.Von Neumann ως παράδειγμα για πληροφορίες από τα βιολογικά δίκτυα και δημιουργίας των πρώτων τεχνητών δικτύων.Το 1949 ο D.Hebb με το βιβλίο του «The organization of behavior » εισάγει τον κανόνα μάθησης του Hebb.Το μοντέλο του Hebb έχει ως κεντρική ιδέα τις συνδέσεις μεταξύ μονάδων του συστήματος ,δηλαδή τους νευρώνες.Ο κανόνας αυτός λέει ότι κάθε φόρα που το δίκτυο χρησιμοποιεί τις νευρωνικές του συνδέσεις ,οι συνδέσεις αυτές ενισχύονται και το δίκτυο πλησιάζει περισσότερο στο να μάθει το πρότυπο το οποίο παρουσιάζεται .Το μοντέλο του αισθητήρα παρουσιάσθηκε για πρώτη φορά το 1957 από τον F.Rosenblatt, ο οποίος αρχικά έφτιαξε το πρώτο δίκτυο με υλικό που μπορούσε να κάνει πολλές και διάφορες εργασίες .Είναι ένα απλό μοντέλο με δύο επίπεδα ,της εισόδου και της εξόδου ,όπου το σήμα προχωρά μονοδρομικά από την είσοδο στην έξοδο .Στο βιβλίο perception του Minsky και Papert γίνεται μια εκτίμηση της χρησιμότητας του προτύπου αυτού ,παρουσιάζοντας όμως και οι περιορισμοί του .

Το 1959 δυο νέα μοντέλα το Adaline και το Madaline ,αναπτύχτηκαν από τους Widrow και Hoff την ίδια περίπου εποχή με την ανάπτυξη του Perceptron. Τα δύο αυτά μοντέλα ήταν από τα πρώτα που χρησιμοποιήθηκαν σε πρακτικά θέματα επιτυχώς .Χρησιμοποιήθηκαν ως φίλτρα για να εξαλείψουν την ηχώ σε τηλεφωνικές γραμμές .Μεγάλη ώθηση στην ανάπτυξη των Νευρωνικών Δικτύων έδωσε το 1982 ο Hopfield.Σε μια εργασία του απέδειξε με αυστηρά μαθηματική απόδειξη πως ένα Νευρωνικό Δίκτυο μπορεί να χρησιμοποιηθεί ως αποθηκευτικός χώρος και επίσης πως μπορεί να επανακτήσει όλη την πληροφορία ενός συστήματος έστω και αν του δοθούν μερικά μόνο τμήματα και όχι ολόκληρο το σύστημα.

Ένα επόμενο βήμα ήταν η πρόοδος στην διαδικασία των δικτύων όταν επινοήθηκε ο κανόνας της διόρθωσης του σφάλματος (error correction learning).Κατά την εκπαίδευση ενός δικτύου ,σε όποια κατάσταση και αν βρίσκεται αυτό σε μία δεδομένη στιγμή,σημασία έχει η απόκλιση που δίνει στην έξοδο του το δίκτυο από την αναμενόμενη τιμή ή τον στόχο που έχουμε θέσει.Αυτή η απόκλιση είναι το σφάλμα που παράγει το δίκτυο την δεδομένη στιγμή .Το σφάλμα ενεργοποίει ένα μηχανισμό ελέγχου ώστε νε επιφέρει μια σειρά από διορθωτικές αλλαγές στα βάρη w των νευρώνων.

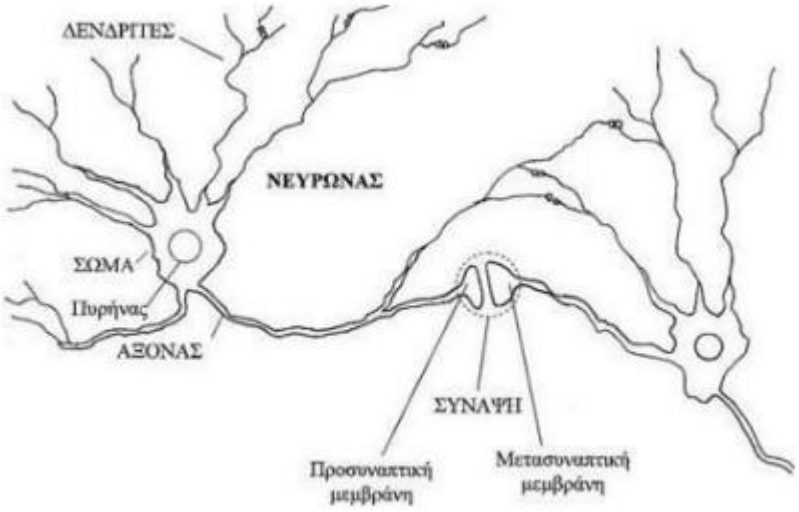
Το 1986 από τους McClelland και Rumelhart δημοσιεύτηκε « το Parallel Distriduted Processing »,στο οποίο παρουσιάζεται η ιδεά πώς ένα Νευρωνικό Δίκτυο μπορεί να χρησιμοποιηθεί ως παράλληλος επεξεργαστής.Το έργο αυτό επιτρέπει την ύπαρξη και άλλων επιπέδων νευρώνων, εκτός από την είσοδο που αποτελούν την εσωτερική δομή του δικτύου.Προτείνουν την μέθοδο της οπισθοδιάδοσης η οποία είναι η πιο χρήσιμη τεχνική εκπαίδευσης των δικτύων. (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

3.15.2 ΒΙΟΛΟΓΙΚΑ ΝΕΩΡΩΝΙΚΑ ΔΙΚΤΥΑ

Το Επιστημονικό έργο στο πεδίο των Τεχνητών Νευρωνικών Δικτύων βασίστηκε στο γεγονός ότι ο ανθρώπινος εγκέφαλος εκτελεί με διαφορετικό τρόπο τους υπολογισμούς σε σχέση με το συμβατικό υπολογιστή.Για την κατανόηση των Τεχνητών Νευρωνικών Δικτύων είναι πολύ σημαντική η αναλογία μεταξύ νευροφυσιολογίας του ανθρώπινου εγκέφαλου και των Τεχνητών Νευρωνικών Δικτύων .Ο νευρώνας είναι ένας εξειδικευμένος τύπος κυττάρου που αποτελεί τη βασική μονάδα των συστημάτων επεξεργασίας πληροφοριών που απαρτίζουν το νευρικό σύστημα του ανθρώπου .Είναι βασικό δομικό στοιχείο εγκεφάλου στον άνθρωπο και στα ζώα.Ο εγκέφαλος ενός νεογέννητου ανθρώπου αποτελείται από περίπου 100 δισεκατομμύρια νευρώνες κάθε ένας από τους οποίους συνδεέται με περίπου 1000 άλλους νερώνες .

Ο Μέσος ανθρώπινος εγκέφαλος εκτιμάται ότι αποτελείται από το 109 νευρώνες που συνδέονται μεταξύ τους με διάφορους τρόπους .Ανατομικά ο νευρώνας αποτελείται από το σώμα (body)που είναι ο πυρήνας του,τους δενδρίτες (denrites) μέσων των οποίων λαμβάνει σήματα από γειτονικούς νευρώνες (σημεία εισόδου)και τον άξονα (axon)που είναι η έξοδος του νεύρωνα και το μέσο συνδέσης του με άλλους νευρώνες .Σε κάθε δενδρίτη υπάρχει ένα απειροελάχιστο κένο που ονομάζεται σύναψη(synapse).Πιο ειδικά:

Οι δενδρίτες είναι οι πύλες εισόδου του νευρώνα κάθως δέχονται ηλεκρικά σήματα από άλλους νευρώνες.Ο άξονας είναι η πύλη εξόδου του νευρώνα .Στέλνει σήματα προς άλλους νευρώνες υπό μορφή ηλεκτρικών παλμών σταθερού πλάτους αλλά μεταβλητής συχνότητας .
Οι συνάψεις είναι τα σημεία που ενώνονται οι διακλαδώσεις του άξονα ενός νευρώνα με τους δενδρίτες άλλων νευρώνων. Είναι κύστες με ηλεκτροχημικό υλικό.Το υλικό αυτό μεταδίδει την ηλεκτρική δραστηριότητα του άξονα-αποστολέα στους δενδρίτες παραλήπτες.Από το πλάτος της σύναψης,την απόσταση της από τον δενδρίτη και από την πυκνότητα του ηλεκροχημικού υλικού εξαρτάται η ευκολία με την οποία η ηλεκτρική δραστηριότητα μεταδίδεται από τον άξονα στο δενδρίτη.Το συνοπτικό βάρος είναι το ποσοστό της ηλεκτρικής δραστηριότητας που μεταδίδεται τελικά στο δενδρίτη.Οι συνάψεις χωρίζονται σε ενισχυτικές και ανασταλτικές ανάλογα με το αν το φορτίο που εκλύεται από την σύναψη ερεθίζει τον νευρώνα να παράγει παλμούς με μεγαλύτερη συχνότητα ή αν τον καταστέλλει εμποδίζοντας τον να παράγει παλμούς αντίστοιχα. Στο σχήμα φαίνεται η αναπαράσταση ενός βιολογικού νευρώνα.

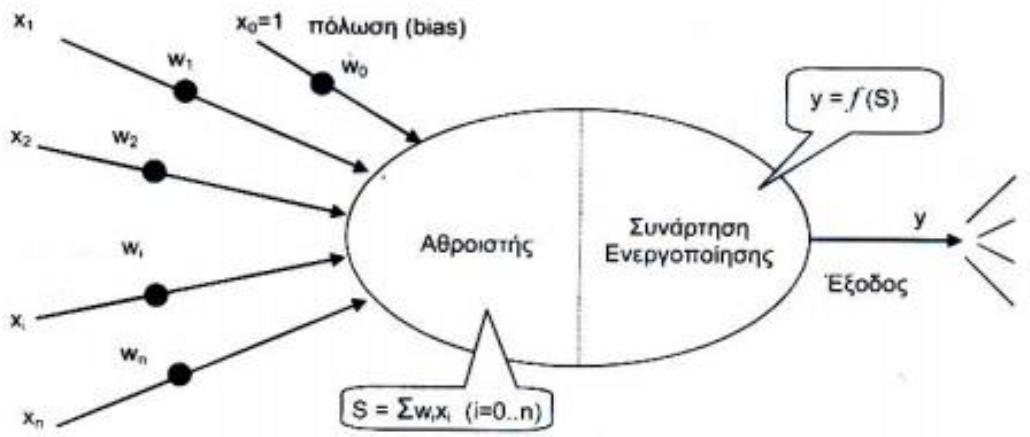


Σχήμα 14:Βιολογικός νευρώνας

Το βασικότερο χαρακτηριστικό του νευρώνα είναι η ευαισθησία του ,δηλαδή η ικανότητα του να αντιδρά σε διάφορα εξωτερικά ερεθίσματα του (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

3.15.3ΤΟ ΜΟΝΤΕΛΟ ΤΟΥ ΤΕΧΝΗΤΟΥ ΝΕΥΡΩΝΑ

Ο τεχνητός νευρώνας είναι ένα υπολογιστικό μοντέλο του οποίου τα μέρη μπορούν να αντιστοιχηθούν με αυτά του βιολογικού νεύρωνα. Ένας τεχνητός νευρώνας δέχεται σήματα εισόδου x1, x2,…,Xn και κάθε τέτοιο σήμα μεταβάλλεται από μια τιμή βάρους wi (weight),ο πόλος της οποίας είναι αντίστοιχος του ρόλου της οποίας είναι αντίστοιχος του ρόλου της σύναψης στο βιολογικό νευρώνα.Η τιμή του βάρους μπορεί να είναι θετική ή αρνητική σε αντιστοίχια με την επιταχυντική ή επιβραδυντική λειτουργία της σύναψης.Το σχήμα παρακάτω παρουσιάζει το μοντέλο του τεχνητού νευρώνα .



Σχήμα 15 :Μοντέλο Τεχνητού Νευρώνα

Το σώμα του τεχνητού νευρώνα χωρίζεται σε δύο μέρη ,τον αθροιστή(sym) ,ο οποίος προσθέτει τα επηρεασμένα από τα βάρη σήματα εισόδου που βγάζει την ποσότητα S και τη συνάρτηση ενεργοποίησης,ένα είδος φίλτρου το οποίο διαμορφώνει την τελική τιμή του σήματος εξόδου y,σε συνάρτηση με την ποσότητα S και την τιμή κατωφλιού της συνάρτησης ενεργοποίησης .Η μοναδικότητα της εξόδου του νευρώνα διευκρινίζεται ότι έχει να κάνει με την τιμή εξόδου και όχι με το πόσες γραμμές –έξοδοι υπάρχουν στο δεξιό μέρος .Ένας νευρώνας μπορεί να έχει πολλές εξόδους ,αλλά όλες θα έχουν την ίδια τιμή.

Η έξοδος του τεχνητού νευρώνα προκύπτει από την εφαρμογή της συνάρτησης ενεργοποίησης στην Συνολική του είσοδο S

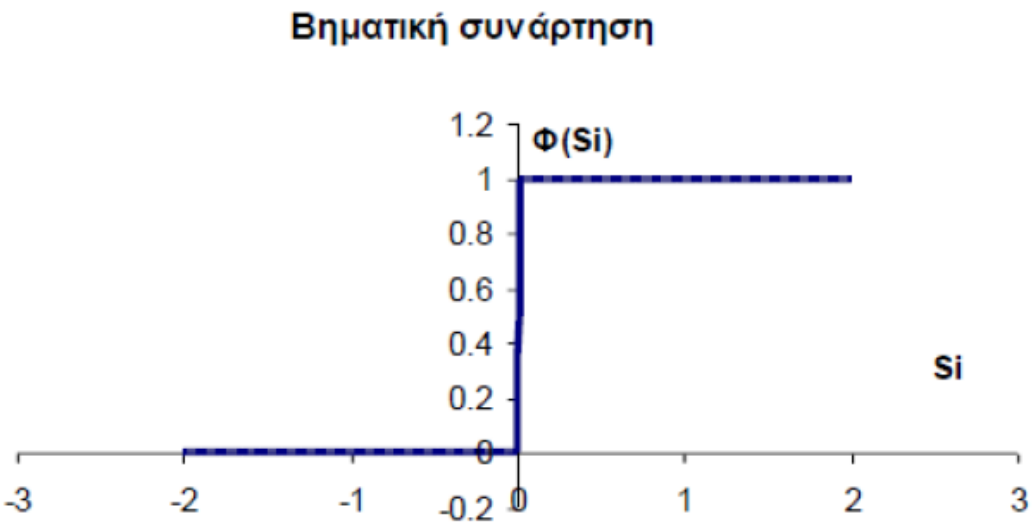
Υπάρχουν διάφορες περιπτώσεις συναρτήσεων ενεργοποίησης :

1. Βηματική συνάρτηση (step fuction) ή συνάρτηση κατωφλίου
2. Συνάρτηση πρόσημου (sign fuction)
3. Σιγμοειδής συνάρτηση (sigmoid ή logistics)
4. Γραμμική συνάρτηση(linear fuction) (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

<u>Βηματική συνάρτηση ενεργοποίησης</u>

Η βηματική συνάρτηση ενεργοποίησης είναι:

$$\phi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

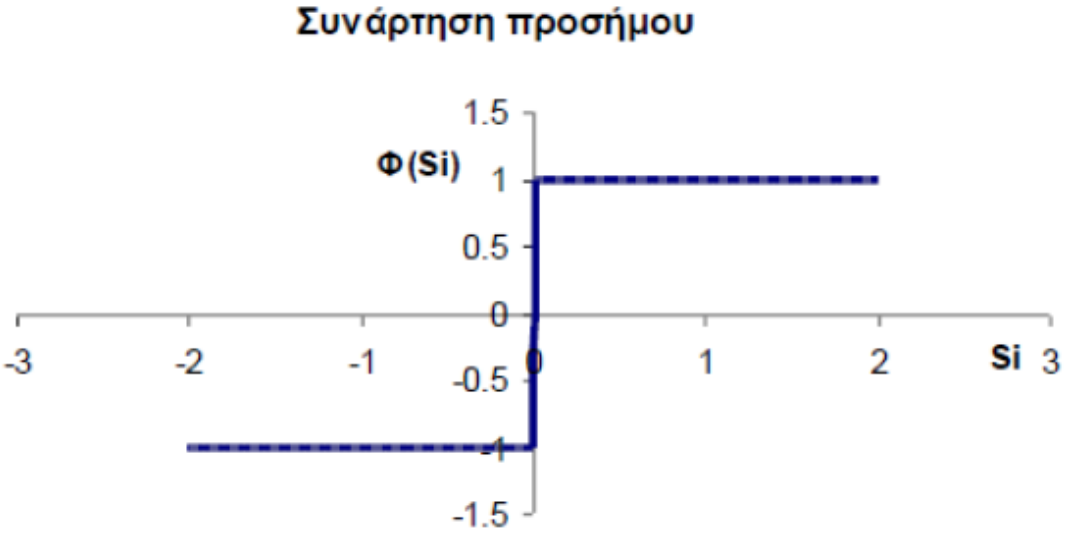


Σχήμα 16: Βηματική συνάρτηση

Συνάρτηση προσήμου

Η συνάρτηση προσήμου είναι:

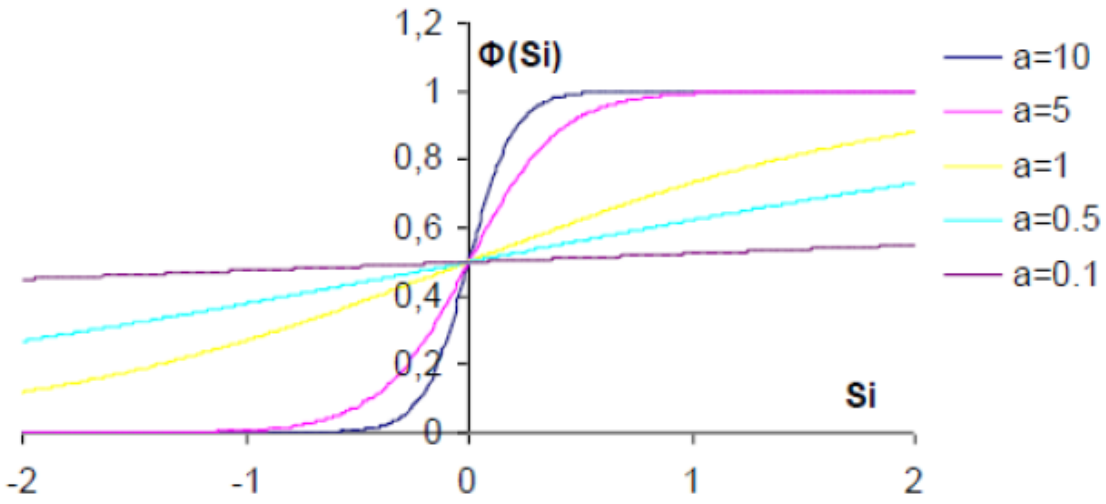
$$\Phi(S) = \begin{cases} 1, \alpha\nu\ S > 0 \\ -1, \alpha\nu\ S \le 0 \end{cases}$$



Σχήμα 17 :Συνάρτηση προσήμου

Σιγμοειδής Συνάρτηση

$$\Phi(S_i) = \frac{1}{1 + e^{-a \cdot S_i}}$$

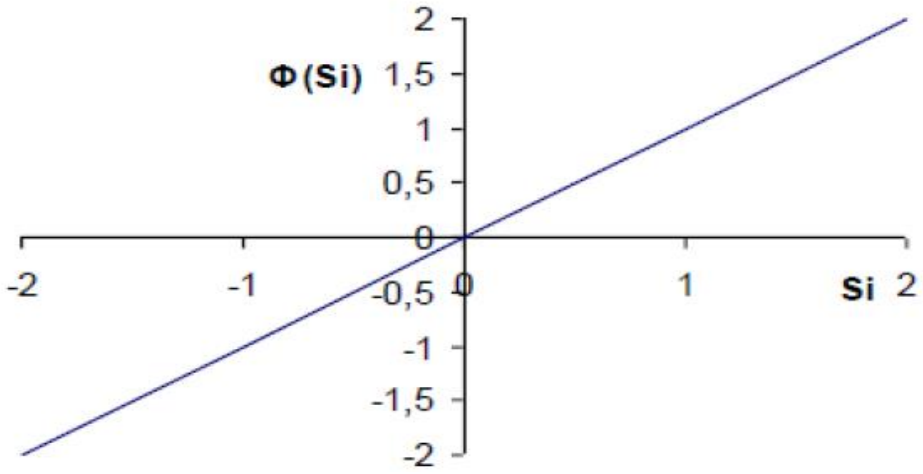


Σχήμα 18 :Σιγμοειδής συνάρτηση

Γραμμική συνάρτηση

$$\Phi(S_i) = \lambda \cdot S_i$$

Γραμμική συνάρτηση ενεργοποίησης



(Κωσταντινος, 2014)

Σχήμα 19 : Γραμμική συνάρτηση ενεργοποίησης

3.15.4 ΕΚΠΑΙΔΕΥΣΗ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Η ικανότητά τους για εκπαίδευση είναι μια από τις βασικές ιδιότητες των Νευρωνικών Δικτύων . Η εκπαίδευση αυτή επιτυγχάνεται μέσω της ανταλλαγής τιμών και βαρών, που αποσκοπεί στη βαθμιαία σύλληψη της πληροφορίας η οποία στη συνέχεια θα είναι διαθέσιμη προς ανάκτηση. Υπάρχουν πολλοί αλγόριθμοι που η εφαρμογή τους έχει στόχο την προσαρμογή των τιμών των βαρών ενός Τεχνητού Νευρωνικού Δικτύου.Οι μέθοδοι μάθησης χωρίζονται σε δύο κατηγορίες : τη μάθηση με επίβλεψη (supervised learning) και τη μάθηση χωρίς επίβλεψη (unsupervised learning).

Μάθηση με επίβλεψη: Η μάθηση αυτή είναι μια διαδικασία η οποία συνδυάζει έναν εξωτερικό εκπαιδευτή και τη συνολική ή γενικευμένη πληροφορία. Κάποιες από τις μεθόδους οι οποίες συγκαταλέγονται σε αυτή την κατηγορία είναι η μάθηση με διόρθωση σφάλματος, η στοχαστική μάθηση. Παραδείγματα τα οποία αντιπροσωπεύουν την μάθηση με επίβλεψη συμπεριλαμβάνουν αποφάσεις για το πότε θα πρέπει να σταματήσει η διαδικασία εκπαίδευσης, αποφάσεις αναφορικά με τη συχνότητα παρουσίασης στο δίκτυο τα πρότυπα εκπαίδευσης και η παρουσίαση προόδου του δικτύου. Η μάθηση με επίβλεψη χωρίζεται σε δύο ακόμη κατηγορίες: στη δομική (structural) και στην προσωρινή(temporal) εκμάθηση. Οι αλγόριθμοι στην πρώτη κατηγορία, χρησιμοποιούνται για την εύρεση της βέλτιστης σχέσης μεταξύ εισόδων και εξόδων για κάθε ξεχωριστό ζευγάρι προτύπων. Παραδείγματα της δομικής εκμάθησης αποτελούν η αναγνώριση και η κατηγοριοποίηση προτύπων, ενώ παραδείγματα της προσωρινής εκμάθησης είναι η πρόβλεψη και ο έλεγχος.

Μάθηση χωρίς επίβλεψη: Οι αλγόριθμοι της εν λόγω μάθησης αναφέρονται ως αυτό-οργανώμενοι (self-organized) και είναι διαδικασίες οι οποίες δεν απαιτούν να είναι παρών ένας «εξωτερικός» δάσκαλος ή επιβλέπων. Βασίζονται, μάλιστα, μόνο σε τοπική πληροφορία καθ' όλη τη διάρκεια της εκπαίδευσης του Τεχνητού Νευρωνικού Δικτύου. Οι συγκεκριμένοι αλγόριθμοι οργανώνουν τα δεδομένα και ανακαλύπτουν τις σημαντικές συλλογικές ιδιότητες. Οι αλγόριθμοι εκπαίδευσης χωρίς επίβλεψη είναι ο αλγόριθμος Hebbian, ο διαφορικός αλγόριθμος Hebbian και ο Min-Max αλγόριθμος.

Κατά κύριο λόγο οι περισσότερες διαδικασίες εκπαίδευσης είναι off line. off line εκπαίδευση είναι όταν χρησιμοποιείται όλο το δείγμα προτύπων για την τροποποίηση των τιμών των βαρών, πριν την τελική χρήση του δικτύου ως εφαρμογή. Οι αλγόριθμοι εκπαίδευσης off line έχουν την απαίτηση να βρίσκονται στην εκπαίδευση του δικτύου παρόντα όλα τα πρότυπα. Το γεγονός αυτό αποκλείει την πιθανότητα εισαγωγής νέων πληροφοριών μέσω νέων προτύπων. Υπάρχουν και Τεχνητά Νευρωνικά Δίκτυα τα οποία δεν αποκλείουν την εισαγωγή νέας πληροφορίας, και μετά την τελική τους μοντελοποίηση. Αν παρουσιαστεί ανάγκη εισαγωγής νέου προτύπου στο δίκτυο, μπορεί να γίνει απευθείας χωρίς τον κίνδυνο να χαθεί κανένα μέρος της αρχικής πληροφορίας. Το πλεονέκτημα των δικτύων που χρησιμοποιούν off line είναι η διαδικασία εκπαίδευσης που επικεντρώνεται κυρίως στη δυνατότητα να δίνουν καλύτερες λύσεις σε δύσκολα προβλήματα. [6]

3.15.5 ΕΦΑΡΜΟΓΈΣ ΤΩΝ ΝΕΥΡΩΝΙΚΏΝ ΔΙΚΤΎΩΝ

Τα Τεχνήτα Νευρωνικά Δίκτυα χρησιμοποιούνται τα τελευταία χρόνια σε διάφορους τομείς λόγω της υπολογιστικής τους ταχύτητας της ικανότητας αντιμετώπισης πολύπλοκων μη γραμμικών λειτουργίων και της ικανότητας τους να αναγνωρίζουν τις σχέσεις μεταξύ ποσοτήτων οι οποίες είναι δύσκολο να μοντελοποιηθούν .Η χρήση των Τεχνητών Νευρωνικών Δικτύων σε διάφορες εφαρμογές προσφέρει ευκολία υλοποίησης ,σχετικά αξιόπιστη λειτουργία και άμεση απόκριση κατά την φάση πραγματικής λειτουργίας εφόσον το Νευρωνικό Δίκτυο υλοποιηθεί σε υλικό (hardware).Οι εφαρμογές αυτές περιλαμβάνουν αναγνώριση προτύπων ,υπολογισμό συναρτήσεων,βελτιστοποίηση,πρόβλεψη,αυτόματο έλενχο και αλλα θέματα. Κάποιες συγκεκριμένες περιοχές εφαρμογών είναι οι παρακάτω :

Βιολογία :βοήθεια στην κατανόηση της λειτουργίας του εγκεφάλου και δημιουργία μοντέλων για την όραση.

Γεωλογία :ανάλυση πιθανότητας ύπαρξης πετρελαίου σε γεωλογικά πετρώματα όπως και ανάλυση πετρωμάτων σε ορυχεία.

Βιομηχανία :Αυτοματοποίηση ρομπότ και συστημάτων έλεγχου,εφαρμογές σε οχήματα, έλεγχος χημικών διεργασιών,έλεγχος στη γραμμή παραγώγης, βιομηχανικός έλεγχος ποιότητας ,επιλόγη ανταλλακτικών κατά τη συναρμολόγηση, ρύθμιση ηλεκρικού φορτίου.
Επεξεργασία σημάτων:Αφαίρεση θορύβου από τηλεφωνική γραμμή, μοντελοποίηση σήματος .

Υπολογιστές: Στους υπολογιστές χρησιμοποιείται για την αναγνώριση προτύπων πχ.αυτόματη αναγνώριση χειρόγραφων χαρακτήρων όπως και για αυτομάτη αναγνώριση προσώπου και φώνης .

Περιβάλλον:Στο περιβάλλον οι εφαρμογές που έχει είναι για την πρόγνωση του καιρού και για την ανάλυση καιρικών συνθηκών.

Αεροπλοία: Χρησιμοποιείται για την δημιουργία αυτόματων πιλότων και προγραμμάτων προσομοίωσης πτήσης και για συστήματα έλεγχου πτήσης .(ΛΑΟΤΡΑ, 2013)

Ιατρική διάγνωση: Ένα ευρύ φάσμα ιατρικά συσχετιζόμενων ενδείξεων, όπως ο συνδυασμός της καρδιακής συχνότητας, τα επίπεδα των διαφόρων ουσιών στο αίμα, ο ρυθμός της αναπνοής μπορούν να παρακολουθηθούν. Η εκδήλωση μιας συγκεκριμένης ιατρικής κατάστασης, γίνεται να συσχετιστεί με ένα πολύπλοκο συνδυασμό μεταβολών σε ένα υποσύνολο μεταβλητών που παρακολουθούνται. Τα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί για την αναγνώριση αυτού του προτύπου πρόβλεψης, ώστε να χορηγηθεί η κατάλληλη θεραπεία.

Χρηματιστηριακές προβλέψεις: Τα νευρωνικά δίκτυα χρησιμοποιούνται από πολλούς τεχνικούς αναλυτές, για να κάνουν προβλέψεις σχετικά με τις τιμές των μετοχών, βασιζόμενοι σε ένα μεγάλο αριθμό παραγόντων, όπως τις προηγούμενες επιδόσεις άλλων αποθεμάτων και διαφόρων οικονομικών δεικτών.

Πιστωτική ανάθεση: Μια ποικιλία από κομμάτια πληροφοριών, τα οποία είναι συνήθως γνωστά για ένα απαιτούμενο δάνειο. Για παράδειγμα, η ηλικία του αιτούντος, η εκπαίδευση, το επάγγελμα και πολλά άλλα στοιχεία που μπορεί να είναι διαθέσιμα. Μετά την εκπαίδευση ενός νευρωνικού δικτύου σε ιστορικά δεδομένα η ανάλυση μπορεί να εκτοπίσει τα πιο κατάλληλα και σχετικά χαρακτηριστικά και να τα χρησιμοποιήσει για την ταξινόμηση των αιτούντων ως χαμηλού ή υψηλού κινδύνου.

Συστήματα διαχείρισης κινητήρα:Οι αισθητήρες ενός κινητήρα χρησιμοποιούν τα νευρωνικά δίκτυα για την ανάλυση των εισροών που δέχονται. Προκειμένου να επιτευχθεί ένας συγκεκριμένος στόχος το νευρωνικό δίκτυο ελέγχει μια ποικιλία παραμέτρων με τις οποίες λειτουργεί ο κινητήρας.Για παράδειγμα επιχειρείται η ελαχιστοποίηση της κατανάλωσης των καυσίμων μέσω του δικτύου αυτού.

<u>Παρατήρηση</u>

Οι παραπάνω εφαρμογές απαιτούν σε μικρότερο ή μεγαλύτερο βαθμό τη διαδικασία της μηχανικής μάθησης.Τα νευρωνικά δίκτυα δεν αποτελούν το μοναδικό τρόπο αντιμετώπισης

των παραπάνω εφαρμογών.Όμως η χρήση τους μας προσφέρει γενικά ευκολία υλοποίησης ,σχετικά αξιόπιστη λειτουργία ,και ακαριαία απόκριση κατά την φάση πραγματικής λειτουργιάς εφόσον το νευρωνικό δίκτυο υλοποιηθεί σε hardware. [7]

3.16 ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Η μέθοδος των δεντρογραμμάτων είναι εδώ και καιρό ευρέως δημοφιλής για τα μελοντικά μοντέλα.Η δημοφιλία τους οφείλεται,σε μεγάλα κομμάτια ,στο γεγονός ότι ήταν απλοίκα και στην προφητική ικανότητα που είχαν μαζί με τον μικρό αριθμό των υποθέσεων.

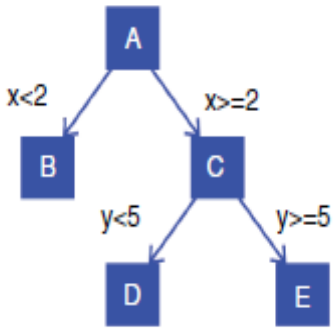
Η μέθοδος των δεντροδιαγραμματών του μοντελισμού υπάρχει από την δεκαετία του ’60 συμπεριλαμβάνοντας δημοφιλής μεθόδους όπως Cart,C4.5, CHAID ως κοινές εφαρμογές.Τα δέντρα απόφασης είχαν πρωτοπαρουσιαστεί από τον Leo Breiman το 1984 στο βιβλίο του με τίτλο Classification and Regression trees.Ο αρχικός αλγόριθμος CART έχει βελτιωθεί και επεκταθεί με πολλούς τρόπους κατά την διάρκεια των χρόνων αλλά και οι συνολικές αρχές παραμένουν.

Ξεκινώντας με ένα σύνολο από εκπαιδεύτικά δεδομένα και παρατηρώντας την εισαγωγή μεταβλητών και αντίστοιχων επιπέδων ώστε να προσπδιορίσουμε και τα πιθανά διαιρεμένα σημεία .Το κάθε διαιρεμένο σημείο είναι το καλύτερο .Το καλύτερο σε αυτήν την περίπτωση είναι να δημιουργηθεί το πιο καθαρό σύνολο από τις υποκατηγορίες .Η διαδικασία τότε επαναλαμβάνεται για κάθε υποκατηγορία μέχρι να συναντήσουμε έναν τερματικό όριο .Η κυριότερη διάκριση μεταξύ των δέντρων απόφασης παραχώρουν μια κατηγορία από περιορισμένη ή ονομαστική λίστα επιλογών όπως είναι το γένος ή το χρώμα .Σε αντίθεση ,τα δέντρα παλινδρόμησης προβλέπουν μια συνεχόμενη αξία για κάθε χώρα όπως είναι το εισόδημα ή η ηλικία .

Εξετάζοντας το πρόβλημα της προσπάθειας να προβλέψουμε το γένος των μαθητών από ένα γενικό εκλετκτικό πρόγραμμα μαθημάτων στο πανεπιστήμιο.

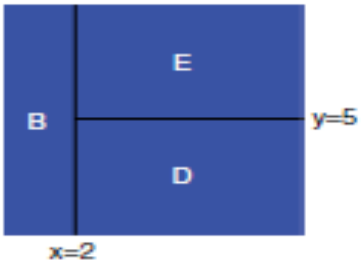
---

[7] (57https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF)



Σχήμα 20 :Βασικό δέντρο απόφασης

Μπορείτε να δημιουργείσεται ένα δέντρο απόφασης το οποίο προβλέπει το γένος από των διαχωρισμό ενός αντιπροσωπευτικού πληθυσμού από το ύψος ,το βάρος,την ημερομηνία γέννησης,το επίθετο ή το όνομα.Παρόλα αυτά όμως το ύψος και το βάρος τα οποία πιθανότατα θα ήταν δυνατόν να προσδιοριστεί το γένος σε κάποιες περιπτώσεις (οι πολύ ψήλοι πιθανότατα θα είναι αρσενικά ενώ οι πολύ κοντοί πιθανότατα θα είναι θηλυκά).Ακόμα πιο προβλεπόμενα ακόμη θα ήταν τα όνοματα διότι υπάρχουν λιγότερα σε αριθμό ονόματα από ότι ύψη ή βάρη που είναι γενετικά ουδέτερα .Μια μεταβλητή που είναι δύσκολη να προσδιοριστεί αλλά μελλοντικά θα γίνει εφικτό να προσδιοριστεί είναι το μήκος των μαλλιών.Αυτό το παράδειγμα δείχνει ότι με τις σωστές ιδιότητες είναι συχνά πιθανόν να αγγίξει τον τέλειο διαχωρισμό από μια ενιαία μεταβλητή. Στα περισσότερα προβλήματα του πραγματικού κόσμου αυτή η ενιαία ‘’μαγική’’ μεταβλητή δεν υπάρχει γιατί είναι πολύ δαπανηρή για να συλλεχθεί ή είναι κάτι το οποίο δεν το γνώριζαν εκ των προτέρων για τον σκοπό αυτό .Τα δέντρα απόφασης και παλινδρόμησης τα οποία ήταν τα γνήσια ,αυθεντικά δεδομένα (regionA) αρχικά είχαν διαχωρίστει σε μεταβλήτες y από το οποίο ήταν λιγότερο από 2 (regionB)ή μεγαλύτερα ή ίσα με 2 (region C) region C είναι πλέον πιθανόν να έχει διαχωρίστει σε 2 περιοχες $D(y < 5)$ και την περιοχή $E(y >= 5)$.Αυτές οι παρόμοιες περιοχές παρουσιάζονται στην εικόνα ως διαχωρισμένες δυνάμεις σε κάθε περιοχή .Αυτές οι περιοχές δείχνουν μόνο τα φύλλα του δέντρου (κάθε φύλλο είναι ένας κόμβος ο οποίος δεν έχει πολλά διαχωρίσματα).Τα φύλλα τα οποία διαχωρίζονται ονομάζονται διακλαδώσεις .Οι διακλάδωσεις στην εικόνα είναι η A και η C .Η δημιουργία ενός δέντρου γίνεται σε 2 φάσεις. Αρχικά,ένα μεγάλο δέντρο δημιουργείται από την επανάληψη των δεδομένων για το συλλογικό εισόδημα χρησιμοποιώντας ένα μετρικό για να προσδιοριστεί –υπολογιστεί το καλύτερο .



Σχήμα 21 : Βασικό δέντρο απόφασης εμφανίζεται ως διχοτομική περιοχή

Η ανάπτυξη συνεχίζεται μέχρι κάποιο τερματικό σημείο να φτάσει σε κάποια συνθήκη η οποία θα συμπεριλαμβάνει ανεπαρκή δεδόμενα για splits δεν μπορούν να βελτιώσουν τα μοντέλα που ταιριάζουν ή το δέντρο έχει φτάσει σε ένα συγκεκριμένο βάθος.Αφού έχει δημιουργηθεί το δέντρο ,είναι συχνά overfit (το οποίο κάνει το δέντο φτωχό για να γίνει κάθε πρόβλεψη).Αυτό ταιριάζει με τα δεδομένα εισόδου παρά πολύ καλά ,ταιριάζει με όλες τις μικρές διακυμάνσεις που είναι παρούσες επί των δεδομένων εισόδου ,αλλα όχι στον γενικό πληθυσμό.Έτσι το πολύ μεγάλο αυτό δέντρο στην συνέχεια επανέρχεται μέχρι να είναι σε γενικές γραμμές προγνωστικό ,όχι μόνο για το συγκεκριμένο σύνολο δεδόμενων εισόδου.

Η περικοπή είναι ιδαίτερα αποτελεσματική εάν ένα δεύτερο σύνολο δεδομένων χρησιμοποιηθεί ώστε να καθορίσει πόσο καλά αποδίδει το δέντρο στα γενικά δεδομένα .Αυτό έρχεται στο κόστος παίρνωντας κάποια δεδομένα που θα έχουν χρησιμοποιηθεί ώστε να μεγαλώσει το δέντρο .Η μέθοδος C4.5 επιχειρεί να περικόψει με βάση τα αναμενόμενα ποσοστά σφάλματος αντί να χρησιμοποιηθεί ένα ξεχωριστό σύνολο δεδομένων.Υπάρχει ένας μεγάλος αριθμός μεθόδων για τον προσδιορισμό του ελέγχου καλής προσαρμόγης της διάστασης .(Dean, 2014)

3.17 ΔΙΚΤΥΑ BAYES

Ο Thomas Bayes ήταν προτεστάντης υπουργός που έζησε στην Αγγλία το δέκατο όγδοο αιώνα. Θεωρήθηκε ένας ερασιτέχνης μαθηματικός και ανέπτυξε το δικό του τώρα διάσημο θεώρημα , καθώς προσπαθούσε να αποδείξει την ύπαρξη του Θεού μέσα από την αντίστροφη πιθανότητα των αιτίων . Το πεδίο της πιθανότητα ήταν πολύ πρωίμο εκείνη την στιγμή, ως επί το πλείστον περιοριζόταν σε παιχνίδια ευκαιρίας όπως το πόκερ, έτσι ο Bayes ανέπτυξε ένα πείραμα σκέψης (η προσομοίωσεις του 1700). Το πείραμα ξεκίνησε με μια μπάλα τυχαία τοποθετημένη σε ένα τέλειο επίπεδο τραπέζι, απαρατήρητο από τον Bayes. Στη συνέχεια, μια άλλη μπάλα ήταν τυχαία πάνω στο τραπέζι, και ένας βοηθός θα απαντούσε στην ερώτηση: Είναι η πρώτη μπάλα προς τα δεξιά ή προς αριστερά της δεύτερης μπάλας; Διαδοχικές μπάλες θα μπορούσαν να ήταν τοποθετημένες τυχαία πάνω στο τραπέζι. Με περισσότερες μπάλες στο τραπέζι ,ο Bayes θα μπορούσε να απεικονίσει την αύξηση της

ακρίβειας της θέσης της πρώτης μπάλας όπως έχουμε πει σε σχετική θέση από κάθε νέα μπάλα.

Οι Ιδέες του Bayes έχουν σε μεγάλο βαθμό μέχρι πρόσφατα αγνοηθεί από το Στατιστικό πεδίο. Καθώς η μέθοδος bayes έχει κάποια μειονεκτήματα, οι άνθρωποι από την φύση τους ενεργούν με τη χρήση τον Bayesian Αρχών.Για να δείτε το Θεώρημα του Bayes σε δράση, να εξετάστε το έργο της οδήγησης ενός αυτοκινήτου. Ποιός είναι πιο πιθανό να εμπλακεί σε ένα αυτοκινητιστικό ατύχημα, ένα άτομο ηλικίας μεταξύ των 16 και 24 ή 25 και 69; Αυτό δεν είναι μια ερώτηση παγίδα.Πιθανώς δεν έχετε μαντέψει, οι έφηβοι είναι χειρότεροι οδηγοί και πιο πιθανόν να εμπλακούν σε ένα ατύχημα. Στην πραγματικότητα αυτό είναι ένα αποτέλεσμα σύγχυσης.Οι έφηβοι έχουν ένα φυσικό πλεονέκτημα καλύτερη όραση, ταχύτερο χρόνο απόκρισης, και καλύτερη ακοή ακόμα τα δεδομένα δείχνουν ότι οι έφηβοι είναι τέσσερις φορές πιο πιθανό να εμπλακούν σε ένα αυτοκινητικό ατύχημα από ό, τι οι ενήλικες. Επιπλέον, ενώ οι έφηβοι αποτελούν μόνο 10% του πληθυσμού, έσπασε το παγκόσμιο ρεκορ συμμετόχων σε ποσοστό 12% του συνόλου των θανατηφόρων ατυχημάτων.Όχι μόνο οτι οι έφηβοι είναι στα περισσότερα ατυχήματα,αλλα τα ατυχήματα που αφορούν τους εφήβους έχουν την τάση να είναι πιο σοβαρά.Μία αιτιολογία που θα μπορούσε κανείς να σκεφτεί για να εξηγήσει αυτή την διαπίστωση ίσως είναι ότι οι έφηβοι είναι πιο πιθανό να εμπλακούν σε επικίνδυνη συμπεριφορά, όπως μιλώντας κατά την οδήγηση. Αυτό στην πραγματικότητα δεν είναι αλήθεια. Σε μια μελέτη από το State Farm Insurance,65% των γονέων παραδέχθηκαν να μιλάνε στο τηλέφωνο κατά την οδήγηση και μόνο ένας μικρός αριθμός εφήβων έχουν την ίδια συμπεριφορά. (dean)Η Bayesian κατηγοριοποίηση αποτελεί μία κατηγορία μεθόδων της κατηγοριοποίησης και βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Πραγματοποιείται μια πιθανοτική πρόβλεψη, προβλέπει την πιθανότητα ένα δείγμα Χ να ανήκει σε κάποια κατηγορία. Για την επαγωγική κατασκευή ταξινομητών η Μάθηση κατά Bayes αποτελεί μια ιδιαίτερα δημοφιλή προσέγγιση, αφενός διότι εκπορεύεται από τον οικείο χώρο του Πιθανοτικού Λογισμού, αφετέρου διότι έχει επιδείξει σημαντικά αποτελέσματα σε ένα ευρύτατο φάσμα εφαρμογών . Η λειτουργία αυτής της κατηγορίας αλγορίθμων στηρίζεται στην υπόθεση ότι η υπό εκμάθηση έννοια σχετίζεται άμεσα με την κατανομή των πιθανοτήτων που παρουσιάζουν τα στιγμιότυπα του προβλήματος αναφορικά με την κλάση στην οποία ανήκουν.

Τα πιο σημαντικά πλεονεκτήματα της προσέγγισης αυτής είναι :

Η δυνατότητα αξιολόγησης των υποθέσεων στις οποίες καταλήγει ο αλγόριθμος μάθησης, μέσω της συσχέτισης ενός βαθμού εμπιστοσύνης της ορθότητάς τους, που αντιστοιχεί στην υπολογισθείσα πιθανότητα να είναι συνεπείς με την πλειοψηφία των παρατηρούμενων δεδομένων. Το χαρακτηριστικό αυτό συνεισφέρει στην παραγωγή εύρωστων μοντέλων, που εξασφαλίζουν ότι η αλήθεια μιας υπόθεσης δεν αμφισβητείται από μεμονωμένες περιπτώσεις στιγμιοτύπων για τις οποίες η υπόθεση κρίνεται ασυνεπής.

Τη συμβολή της στη βαθύτερη κατανόηση και ανάλυση αλγορίθμων μάθησης οι οποίοι δε χειρίζονται απ' ευθείας πιθανότητες. Ένα χαρακτηριστικό παράδειγμα της ιδιότητας αυτής αποτελεί η μελέτη της επαγωγικής προδιάθεσης (inductive bias) ενός αλγορίθμου, του συνόλου των υποθέσεων δηλαδή στις οποίες στηρίζεται ο αλγόριθμος, ώστε να παράγει ένα μοντέλο ικανό να γενικεύει τις υποθέσεις στις οποίες κατέληξε κατά το χειρισμό άγνωστων στιγμιότυπων.

Την παροχή ενός μέτρου σύγκρισης έναντι άλλων μεθόδων Μ.Μ., καθώς οι αλγόριθμοι της κατηγορίας αυτής εγγυώνται τη βέλτιστη επίλυση ενός προβλήματος, δεδομένου ενός συνόλου υποθέσεων που απλοποιούν την κατασκευή του μοντέλου. (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

Στόχος κατηγοριοποίησης :

Δοθέντος ένος προτύπου $X = [x(1), x(2) \dots x(d)]$ € Rd, θα πρέπει να ταξινομηθεί σε μία από τις k κατηγορίες $C1, C2, \dots . Ck$

Θεώρημα Bayes

Δοθέντος του X € Rd και των k κλάσεων ,το θεώρημα Bayes δηλώνει πως :

$$P(Ci|x)p(x) = p(x|ci)P(ci)$$

$P(Ci)$ είναι εκ των προτέρων πιθανότητα της κλάσης Ci

$P(x)$ είναι η συνάρτηση πυκνότητας πιθανότητας του γεγονότος x.

$P(x|Ci)$ είναι η υπο συνθήκη πιθανότητας του γεγονότος χ δοθέντος της κλάσης Ci.

$P(Ci|x)$ είναι η εκ των υστέρων πιθανότητα της κλάσης Ci δοθέντος του x (P.N. Tan, 2006)

3.17.1 ΑΠΛΟΙΚΟΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ ΤΟΥ BAYES (NAIVE BAYES)

Ο Naïve Bayesian είναι ο απλούστερος Bayesian κατηγοριοποιητής, υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των υπόλοιπων γνωρισμάτων. Ο λόγος που γίνεται αυτό είναι για να αποφεύγονται οι πολύπλοκοι υπολογισμοί κατά τη συνθήκη ανεξαρτησίας της κατηγορίας [8] Τα χαρακτηριστικά μιας κλάση που ανήκει το πρότυπο .Κάθε ένα χαρακτηριστικό ακολουθεί προτύπου $x = [x1, x2, \dots . Xd]T$ είναι στατικά ανεξάρτητα ,δηλαδή δεν υπάρχει εξάρτηση μεταξύ των χαρακτηριστικών σε σχέση με την μονοδιάστατη κανονική κατανομή. Αυτές οι υποθέσεις δεν ισχύουν πάντα.

Συμβολισμοί : Σύνολο δεδομένων –προτύπων :D={xi,i=1..N} όπου κάθε πρότυπο xi παριστάνει d μετρήσεις από d χαρακτηριστικά .A1,A2…Ad και ταξινομείται σε μια από τις κλάσεις .

8

(https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B7%CE%B3%CE%BF%CF%81%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7)

Υποτίθεται ότι υπάρχουν m κλάσεις $C1, C2$

Χαρακτηριστικά των Naïve Bayes κατηγοριοποιητών :

- Εύρωστοι σε απομονωμένα σημεία θορύβου
- Μπορούν να διαχειριστούν ελλιπής τιμές
- Εύρωστοι στην ύπαρξη μη σχετικών χαρακτηριστικών .

Αν Xi είναι ένα τέτοιο χαρακτηριστικό ,τότε η $P(Xi|Ci)$ κατανέμεται σχεδόν ομοιόμορφα ,χώρις να έχει επίδραση στο συνολικό υπολογιμό της εκ των υστέρων πιθανότητας.

Τα συχετιζόμενα χαρακτηριστικά μπορούν να μειώσουν την απόδοση των απλοικών bayes κατηγοριοποιητών ,επειδή η υπόθεση της υπό συνθήκη ανεξαρτησίας δεν ισχύει πλέον για αυτά. (P.N. Tan, 2006)

3.18 ΤΜΗΜΑΤΟΠΟΙΗΣΗ

Η τμηματοποίηση είναι η ομαδοποίηση αντικειμένων σε παρόμοιων κατηγορίες αντικειμένων . Ένα τμήμα (cluster) είναι η συλλογή από πανομοιότυπα αντικείμενα που διαφέρουν από τα υπόλοιπα αντικείμενα των υπολοίπων τμημάτων. Δεδομένου ενός συνόλου παραδειγμάτων, η διαδικασία τμηματοποίησης περιλαμβάνει την κατάτμηση των παραδειγμάτων σε υποσύνολα/τμήματα. Στόχος είναι να επιτευχθεί εντός των τμημάτων υψηλή ομοιότητα μεταξύ των αντικειμένων και χαμηλή ομοιότητα μεταξύ των αντικειμένων που ανήκουν σε διαφορετικά τμήματα. Επίσης η τμηματοποίηση είναι γνωστή και σαν ανάλυση τμημάτων (clustering analysis) στη Στατιστική, σαν τμηματοποίηση πελατών (customer clustering segmentation) σαν διαχείριση των συσχετίσεων των πελατών (customer relationship management) στο marketing και σαν μη επιβλεπόμενη μάθηση (unsupervised learning) στη Μηχανική Μάθηση. Η συμβατική τμηματοποίηση εστιάζει στην ανάλυση τμημάτων που βασίζονται στην απόσταση. Η έννοια της απόστασης (ή της ομοιότητας) είναι καθοριστική σε αυτό το σημείο. Τα αντικείμενα θεωρείται ότι είναι σημεία σε ένα χώρο με μέτρο την απόσταση. Στην τμηματοποίηση με βάση τις αρχές (conceptual clustering), παράγεται μια συμβολική αναπαράσταση των τμημάτων που προκύπτουν σε συνδυασμό με την κατάτμηση σε τμήματα. Επομένως, μπορούμε να σκεφτούμε ότι κάθε τμήμα είναι μια αρχή (όπως περίπου μια κατηγορία σε μια κατηγοριοποίηση). (ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν., 2012)

3.19 ΑΝΆΛΥΣΗ ΚΑΤΆ ΣΥΣΤΑΔΕΣ

Ομαδοποίηση είναι η διαδικασία οργάνωσης αντικειμένων σε παρόμοιες ομάδες ανακαλύπτοντας τα όρια μεταξύ αυτών των ομάδων αλγοριθμικά χρησιμοποιώντας έναν αριθμό διαφορετικών στατιστικών αλγορίθμων και μεθόδων.Η ανάλυση συστάδων δεν κάνει καμία διάκριση μεταξύ των εξαρτώμενων και ανεξάρτητων μεταβλητών. Εξετάζει ολόκληρο

το σύνολο αλληλεξαρτώμενων σχέσεων για να ανακαλύψουν τις σχέσεις ομοιότητας μεταξύ των αντικειμένων, προκειμένου να προσδιορίσουν τις συστάδες. Ανάλυση συστάδων μπορεί να χρησιμοποιηθεί ως μια διάσταση μεθόδου μείωσης στην οποία ο αριθμός των αντικειμένων ομαδοποιούνται σε ένα σετ συστάδων, και στη συνέχεια, σε ένα περιορισμένο συνολο των μεταβλητών που χρησιμοποιούνται για την προγνωστική μοντελοποίηση. Η πρακτική αυτή βοηθάει στη μείωση θεμάτων που σχετίζονται με την πολυσυγγραμμικότητα. Η Ομαδοποίηση μπορεί να ταξινομηθεί σε τρεις κατηγορίες, χωρίς επίβλεψη,ημιεπιβλέψη, και επίβλεψη.

Η ομαδοποίηση διακρίνεται από τις μεθόδους ταξινόμησης,γιατι αναφέρεται σε ένα γνωστό αριθμό ομάδων και ο λειτουργικός της στόχος είναι να αναθέτει νέες παρατηρήσεις στις υπάρχουσες ομάδες οπότε ανήκει στην μάθηση με επίβλεψη (supervised learning).Στην ανάλυση συστάδων δεδομένου ότι δεν γίνεται καμία υπόθεση σχετικά με τον αριθμό των ομάδων ή την δομή της ομάδας ,είναι μια πρωτόγονη τεχνική και γι αυτό ανήκει στην μάθηση χωρίς επίβλεψη (unsupervised learning)ερευνά τη φυσική ομαδοποίηση των αντικειμένων με βάση μη ταξινομημένα δεδομένα .Η ομαδοποίηση γίνεται βάσει των ομοιοτήτων ή των αποστάσεων (ανομοιότητες).

1.Ομαδοποίηση Χωρίς επίβλεψη .Ο στόχος της χωρίς επίβλεψη ομαδοποίησης είναι η μεγιστοποίηση ομοιότητας του συμπλέγματος και την ελαχιστοποίηση ομοιότητα του συμπλέγματος , δίνεται ένα μέτρο ομοιότητας / ανομοιότητας.Χρησιμοποιεί (π.χ., μια λειτουργία μιας συγκεκριμένης αντικειμενικής συνάρτησης που ελαχιστοποιεί τις εντός αποστάσεις). Χρησιμοποιεί ένα σύνολο δεδομένων που δεν έχει καμία μεταβλητή στόχο. Ο K -means και η ιεραρχική ομαδοποίηση είναι η πιο ευρέως χρησιμοποιούμενες ανεξέλεγκτες τεχνικές ομαδοποίησης του κατακερματισμού. (Dean, 2014)

2. Ημιομαδοποίηση. Εκτός από το μέτρο ομοιότητας που επίσης χρησιμοποιείται σε ομαδοποίηση χωρίς επίβλεψη, η ημι εποπτεύομενη ομαδοποίηση χρησιμοποιεί άλλες κατευθυντήριες / προσαρμογή των στοιχείων του τομέα για τη βελτίωση των αποτελεσμάτων ομαδοποίησης. Αυτές οι πληροφορίες στο πεδίο ορισμού μπορεί να είναι κατά ζεύγη περιορισμοί μεταξύ των παρατηρήσεων ή μεταβλητών-στόχων για ορισμένες από τις παρατηρήσεις.Αυτή η κατευθυντήρια πληροφορία χρησιμοποιείται είτε για τη ρύθμιση της ομοιότητας μέτρησης ή την τροποποίηση της αναζήτησης συστάδων. Εκτός από τους στόχους για μη ελενχόμενη ομαδοποίηση, έχει ως στόχο την απόκτηση υψηλής συνοχής μεταξύ των συστάδων. (Dean, 2014)

3.Εποπτευόμενοι ομαδοποίηση. Ο στόχος των εποπτευόμενης ομαδοποίησης είναι να προσδιορίσει συστάδες που έχουν υψηλές πιθανότητες πυκνότητας με τις επιμέρους κατηγορίες (κατηγορία ομοιόμορφη συστάδων). Χρησιμοποιείται όταν υπάρχει μία μεταβλητή στόχος και ένα σύνολο εκπαίδευσης που περιλαμβάνει μεταβλητές στην ομαδοποίηση. (Dean, 2014)

3.20 ΜΈΤΡΑ ΑΠΌΣΤΑΣΗΣ

Ένας αντικειμενικός τρόπος για να αξιολογήσεις την τμηματοποίηση βασίζεται σε μια ομοιότητα ή ανομοιότητα μεταξύ των αντικειμένων και των συστάδων στο σύνολο δεδομένων. Πολλά πιθανά μέτρα απόστασης μπορούν να χρησιμοποιηθούν για να υπολογιστούν οι συστάδες. Οι πιο κοινές αναφέρονται παρακάτω. Δύο τύποι αποστάσεις μπορούν να μετρηθούν και να προσπελαστούνε απόσταση μεταξύ των αντικειμένων και η απόσταση μεταξύ των συστάδων. Και οι δύο είναι σημαντικές μετρήσεις ανάλογα με την εφαρμογή ομαδοποίησης.

1.Μέτρα Απόστασης μεταξύ αντικειμένων

Χρησιμοποιούνται για να μετρήσουν την απόσταση μεταξύ των αντικειμένων στο σύνολο δεδομένων που χρησιμοποιείται για την ομαδοποίηση.

Ευκλείδεια απόσταση. Είναι ουσιαστικά η γεωμετρική απόσταση μεταξύ των αντικειμένων στον πολυδιάστατο χώρο. Αντιστοιχεί προς το μήκος της συντομότερης διαδρομής μεταξύ δύο αντικειμένων. Χρησιμοποιείται για την απόκτηση σχήματος σφαίρας συστάδων.

City block (Manhattan) απόσταση. Αντιστοιχεί στο άθροισμα αποστάσεων κατά μήκος κάθε διάστασης και είναι λιγότερο ευαίσθητη σε ακραίες τιμές. Χρησιμοποιείται για την απόκτηση σχήματος διαμαντιού συστάδων.

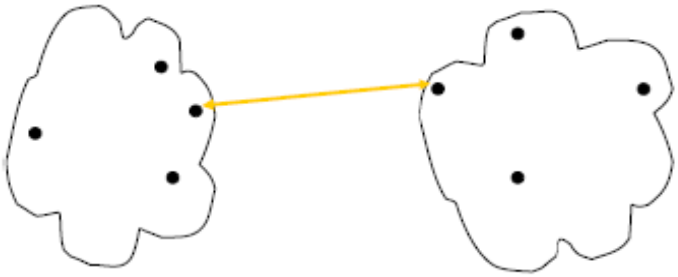
Μέτρο ομοιότητας Συνημιτόνου. Υπολογίζεται με τη μέτρηση
του συνημιτόνου της γωνίας μεταξύ δύο αντικειμένων. Χρησιμοποιείται κυρίως για να υπολογίσουμε την ομοιότητα μεταξύ των δύο συνόλων δεδομένων συναλλαγής. (Dean, 2014)

2.Μέτρα Απόστασης μεταξύ των συστάδων

Χρησιμοποιούνται για να μετρήσουν την απόσταση μεταξύ των συστάδων Τα μέτρα που χρησιμοποιούνται στην ιεραρχική συσταδοποίηση είναι:

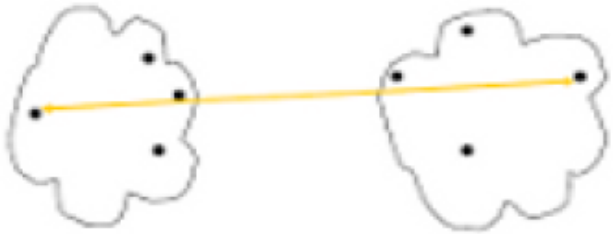
Τεχνική απλής Συνδεσιμότητας

Η εγγύτητα (ή ομοιότητα) μεταξύ συστάδων βασίζεται στα δύο πιο γειτονικά (κοντίνα ή ομοία)σημεία που βρίσκονται σε διαφορέτικες συστάδες.



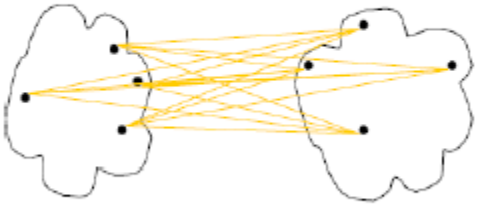
Τεχνική πλήρους Συνδεσιμότητας (Complete Linkage)

Η εγγύτητα(ή ομοιότητα)μεταξύ συστάδων βασίζεται στα δύο πιο μακρυνά(λιγότερο όμοια)σημεία που βρίσκονται σε διαφορετικές συστάδες.



Τεχνική Μέσου όρου ομάδας(Group Average link)

Η εγγύτητα (ή ομοιότητα)μεταξύ δυο συστάδων ορίζεται η μέση εγγύτητα ανά ζεύγη ανάμεσα σε όλα τα ζεύγη σημείων των δύο αυτών.



(P.N. Tan, 2006)

3.21 ΑΞΙΟΛΟΓΗΣΗ ΟΜΑΔΟΠΟΙΗΣΗΣ

Από τότε που η ομαδοποίηση χρησιμοποιείται ως επί τον πλείστον με ένα ανεξέλεγκτο τρόπο πρέπει να υπάρχει μέτρο για την αξιολόγηση της ποιότητας των συστάδων που προβλέπονται από το συγκεκριμένο αλγόριθμο. Αυτά τα κριτήρια αξιολόγησης χρησιμοποιούνται κατά κύριο λόγο για να βρούν τη μεταβλητότητα ή των θόρυβο στις συστάδες, για να βρεθεί βέλτιστος αριθμός των συστάδων για τα δεδομένα, να συγκρίνουν τους αλγορίθμους ομαδοποίησης για την ποιότητα των λύσεων τους καθώς και να συγκρίνουν τα δυο σύνολα αποτελεσμάτων που λαμβάνονται από την ανάλυση ομαδοποίησης.

1.Εσωτερικό κριτήριο αξιολόγησης

Είναι μέρος της ανάλυσης διασποράς και είναι συγκεκριμένη για τη μέθοδο. Υπολογίζεται για τα δεδομένα που χρησιμοποιείται για την ανάλυση συμπλέγματος. Αν και οι

αντικειμενικές συναρτήσεις για όλους τους αλγόριθμους ομαδοποίησης είναι πολύ παρόμοιοι(υψηλή ομοιότητα συμπλέγματος και χαμηλή ομοιότητα συμπλέγματος) το κριτήριο αξιολόγησης που χρησιμοποείται σε διαφορετικούς αλγορίθμους διαφέρει το ένα από το άλλο. Για παράδειγμα, ενώ το άθροισμα τετραγώνων σφάλματος (SSE) μπορεί να χρησιμοποιείται για την αξιολόγηση του k -means αλγόριθμου, δεν μπορεί να χρησιμοποιηθεί για την πυκνότητα που βασίζονται οι αλγόριθμοι ομαδοποίησης.

2. Εξωτερικό κριτήριο αξιολόγησης

Το κριτήριο αυτό μπορεί να ληφθεί χρησιμοποιώντας ένα ξεχωριστό σύνολο δεδομένων που δεν χρησιμοποιήθηκε για την ανάλυση συστάδων. Επίσης χρησιμοποιείται για να μετρήσει το πόσο αντιπροσωπευτικό είναι οι συστάδες σε σχέση με ένα πραγματικό σύμπλεγμα όταν οι ετικέτες κλάσεων δίνονται ή πόσο συνεπείς είναι με σεβασμό στις διαφορετικές συστάδες όταν λαμβάνονται με τη χρήση διαφορετικών παραμέτρων / μεθόδους. Μπορεί να μετρηθεί σε όρους καθαρότητας, εντροπίας, τυχαίου δείκτη, ή F-μέτρου. (Dean, 2014)

3.22 K-MEANS ΑΛΓΟΡΙΘΜΟΣ

Ο MacQeen πρότεινε το όρο K-means ,για να περιγράψει τον αλγόριθμο του , ο οποίος ανήκει στην κατηγορία των διαμεριστικών αλγορίθμων και αναθέτει κάθε στοιχείο στο cluster,που έχει το πλησιέστερο cendroid (κέντρο βάρους)και είναι ένας από τους ευρέως χρησιμοποιημένους και παλιότερους αλγορίθμους. Ο αλγόριθμος διαμέρισης προκαλεί μια διαμέριση του χώρου των δεδομένων ,χωρίς να δημιουργούν πιο πολύπλοκες δομές που περιγράφονται με δενδρογράμματα και κατασκευάζουν μια μοναδική ομαδοποίηση κι όχι μια δομή ομάδων ,όπως ένας ιεραρχικός αλγόριθμος. Κατακερματίζει τα δεδομένα σε k ομάδες, αναθέτοντας σε κάθε αντικείμενο στο πλησιέστερο κέντρο βάρους του συμπλέγματος (η μέση τιμή των μεταβλητών για όλα τα αντικείμενα σε αυτό το συγκεκριμένο σύμπλεγμα) με βάση την απόσταση μέτρο που χρησιμοποιείται. Είναι πιο ισχυρή σε διαφορετικούς τύπους μεταβλητών. Επιπλέον, είναι πιο γρήγορη για μεγάλα σύνολα δεδομένων, τα οποία είναι κοινά σε τμηματοποίηση αλγόριθμοι αυτοί χρησιμοποιούνται πιο πολύ σε περιπτώσεις ,όπου τα δεδομένα είναι πάρα πολλά και η κατασκευή δενδρογραμμάτων είναι αδύνατη. Το κύριο πρόβλημα τους είναι ,η απόφαση για τον αριθμό των τελικών cluster και το κριτήριο που χρησιμοποιείται για την τελική απόφαση, είναι το κριτήριο τετραγωνικού λάθους .

Η κυρία ιδέα του αλγορίθμου ,είναι να καθορίσουμε εμείς ένα συγκεκριμένο αριθμό από κ κέντρα των cluster (cancroids). Όταν θα ξεκινά ο αλγόριθμος ,που φυσικά θα συμβολίζουν και τον αριθμό των τελικών cluster που θα έχουμε και ως έξοδο του αλγορίθμου .Το επόμενο βήμα ,είναι να αναθέσει κάθε δεδομένο στο κοντινότερο του centroid .Όταν ανατεθούν όλα τα δεδομένα ,έχει γίνει το πρώιμο clustering. Στη συνέχεια ,επαναυπολογίζουμε τα centroid

με βάση τα καινούργια cluster που έχουν δημιουργηθεί και τα εναποθέτουμε στο κοντινότερο σε αυτά κέντρο. Με αυτόν τον τρόπο ,έχει δημιουργηθεί ένας βρόχος ,που τερματίζεται όταν πλέον τα κέντρα δεν μετακινούνται από την θέση τους .Ένας από τους βασικούς στόχους του αλγορίθμου είναι να καταφέρει να ελαχιστοποιήσει τη συνάρτηση τετραγώνου λάθους.

Ο βασικός αλγόριθμος για το πρόβλημα k –means λειτουργεί ως εξής:

Λαμβάνει ως παράμετρο εισόδου το πλήθος των επιθυμητών συστάδων k,και τυχαία επιλέγει k αντικείμενα κάθε ένα από τα οποία αρχικά αναπαριστά το κέντρο κάθε συστάδας

Κάθε ένα αντικείμενο ανατίθεται στην συστάδα εκείνη με την οποία είναι πιο όμοιο,βάση της απόστασης μεταξύ του αντικειμένου και του κέντρου της συστάδαςΕνημερώνεται ξανά το νέο της κάθε συστάδας με βάση τα σημεία που αποδίδονται στην συγκεκριμένη συστάδα.Η διαδικασία αυτή συνεχίζεται μέχρι να μην υπάρχει σημείο που να αλλάζει συστάδα ή ισοδύναμα μέχρι τα κέντρα βάρους να μείνουν σταθερά (κριτήριο τερματισμού)

Βήματα αλγόριθμου K means:

1. ***Είσοδος:***k το πλήθος των επιθυμητών συστάδων ,D το σύνολο των δεδομένων που περέχει n αντικείμενα
2. ***Έξοδος:***Ενας σύνολο από k συστάδες
3. ***Επανελαβε***
4. ***Ανάθεσε*** κάθε αντικείμενο στην συστάδα εκείνη στην οποία το αντικείμενο είναι πιο κοντά ,με βάση την απόσταση του αντικείμενου από το κέντρο βάρους της ομάδας.
5. ***Ενημέρωσε τα κέντρα βάρους*** της κάθε συστάδας με βάση τα αντικείμενα που έχουν ανατεθεί σε κάθε συστάδα Μέχρι καμία αλλαγή να λάβει χώρα. (ΓΕΡΑΣΙΜΟΣ, 2013)

3.23 ΙΕΡΑΡΧΙΚΗ ΣΥΣΤΑΔΟΠΟΊΗΣΗ

Η Ιεραρχική ομαδοποίηση παράγει συστάδες που οργανώνονται μέσα σε μια Ιεραρχική δομή. Απεικονίζοντας την Ιεραρχική δομή μπορεί να χρησιμοποιηθεί για την κατανόηση της δομής των συστάδων στο σύνολο δεδομένων. Η τεχνική αυτή απαιτεί μόνο ένα μέτρο της ομοιότητας μεταξύ των αντικειμένων. Δεν απαιτείται να ορίσεις τον αριθμό των συστάδων. Μπορείς να αποκτήσεις οποιοδήποτε αριθμό από συστάδες κόβοντας την ιεραρχική δομή σε σωστό επίπεδο. Υπάρχουν δύο κύριες κατηγορίες για την ιεραρχική ομαδοποίηση :

Διαχωριστική ή διαιρετική (divisive) όπου όλα τα στοιχειά τοποθετούνται σε μια συστάδα και σε κάθε βήμα μια συστάδα διασπάται σε δυο επιμέρους συστάδες μέχρι να απομείνουν μεμονωμένες συστάδες ξεχωριστών σημείων.

Πρέπει να αποφασιστεί ποια συστάδα θα διαχωριστεί και πως θα γίνει αυτός ο διαχωρισμός.

Συσσωρευτική (agglomerative) όπου ξεκινά με τα σημεία ως ξεχωριστές συστάδες σε κάθε βήμα όπου τα πιο κοντινά ζεύγη συστάδων συγχωνεύονται.

Απαιτείται ένας καθορισμός της έννοιας της εγγύτητας των συστάδων .(Dean, 2014)

Συμπέρασμα

Τα συμπεράσματα που προέκυψαν από την έρευνα για την εκπόνηση της πτυχιακής μας εργασίας είναι τα εξής:

Αρχικά η διαχείριση πολύ μεγάλων όγκων δεδομένων είναι ένα από το πιο σημαντικά ανοιχτά ζητήματα στις μέρες μας. Τα δεδομένα που παράγονται από την ολοένα και αυξανόμενη χρήση του διαδικτύου στην καθημερινή μας ζωή αλλά και σε επιχειρηματικά περιβάλλοντα, έχουν αυξηθεί εκθετικά τα τελευταία χρόνια. Για το λόγο αυτό οι τεχνικές που χρησιμοποιούνταν στο παρελθόν δεν επαρκούν πλέον.

Στην συνεχεία τα big data απασχολούν ήδη τις επιχειρήσεις και τις Διευθύνσεις Πληροφορικής τους, αλλά το πρώτο συμπέρασμα είναι ότι πριν ξεκινήσουμε να αναζητούμε την πιο κατάλληλη, για την επιχείρηση μας, λύση θα πρέπει να είναι ξεκάθαρη η επιχειρηματική ανάγκη καθώς και το τι αναμένουμε από την υλοποίησή της.

Σε κάθε περίπτωση, δεν πρέπει να υποτιμήσουμε τις αλλαγές που θα επιφέρει μια τέτοια υλοποίηση, δεδομένου ότι επιδρά άμεσα στη δομή της πληροφορικής (IT organization), με την δημιουργία νέων IT functions όπως storage, BI, data management, predictive analytics, data visualization, που απαιτούν δεξιότητες και χρόνο προσαρμογής στο νέο περιβάλλον.

Ανεξαρτήτως του ορισμού που θα προσδώσει κάποιος στα big data, οι σημαντικότερες έννοιες που συνθέτουν αυτό τον ορισμό ξεκινούν από V, όπως volume: μεγάλοι όγκοι δεδομένων και πιο περίπλοκοι, velocity: ανάγκη για μεγάλες ταχύτητες επεξεργασίας, variety: νέοι τύποι δεδομένων όπως: images, videos, sounds ακόμη και τρισδιάστατα αντικείμενα και σίγουρα veracity: για την πιστότητα και ποιότητα των δεδομένων. Καλούμαστε να επιλέξουμε την πιο αποτελεσματική από πλευράς κόστους λύση με στόχο την αξιοποίηση της πληροφορίας που θα οδηγεί στην έγκαιρη και όσο το δυνατό πιο σωστή λήψη αποφάσεων, δίνοντας το μεγαλύτερο δυνατόν value στην επιχείρηση (μια ακόμη έννοια που ξεκινά από V).

Είναι κοινή αναγνώριση και διαπίστωση όλων ότι η πληροφορία αποτελεί σήμερα ένα από τα σημαντικότερα assets και ότι η αξιοποίηση της οποίας, προσθέτει το απαραίτητο σε αυτήν ανταγωνιστικό πλεονέκτημα που την διαφοροποιεί σε σχέση με τον ανταγωνισμό. Για μια εταιρεια η έγκαιρη αναγνώριση τάσεων – ενδείξεων – συμπεριφορών την κάνει leader των εξελίξεων και όχι απλά follower. Αξιόπιστες τεχνικές λύσεις, όπως αυτές που αναφέραμε παρουσιάστηκαν από τους μεγαλύτερους vendors ήδη υπάρχουν και εξελίσσονται και είναι σίγουρο ότι στο άμεσο μέλλον θα δούμε σημαντικές επενδύσεις στο χώρο των big data και στην Ελλάδα ανεξαρτήτως της οικονομικής κρίσης.

Σίγουρα πάντως θα πρέπει να δούμε και να διαχειριστούμε τα big data, αλλά και τα data γενικότερα, από την οπτική γωνία που τα αναγνώρισαν και τα χαρτογράφησαν στο Word Economic Forum, όπου συγκεκριμένα τονίσθηκε πώς «Personal data is the new oil of the

Internet and the new currency of the digital world» (Τα προσωπικά δεδομένα είναι η νέα «τροφή» για το Διαδίκτυο και η νέα αξία για τον ψηφιακό κόσμο) Meglena Kuneva, European Consumer Commissioner, και «Big Data,Big Impact, declare data a new class of economic asset, like currency or gold» (Μεγάλα δεδομένα, μεγάλος αντίκτυπος, αναγνωρίζοντας τα δεδομένα σαν ένα νέο περιουσιακό στοιχείο, όπως είναι το νόμισμα ή ο χρυσός)!

Πλέον, όντας σε θέση να διαχειριστούμε αποδοτικά πολύ μεγάλο όγκο πολυμεσικών δεδομένων, ανοίγονται πολλοί νέοι ορίζοντες τόσο στον τομέα η δυνατότητα για content based αναζητήσεις. Σε δεύτερο στάδιο, το γεγονός αυτό μπορεί να χρησιμοποιηθεί ως βάση για την εξαγωγή meta-πληροφοριών. Για παράδειγμα μας δίνεται η δυνατότητα να εξάγουμε διάφορα patterns για τις τιμές συγκεκριμένων πεδίων ενός τύπου αντικειμένου και να τα χρησιμοποιήσουμε για την κατηγοριοποίηση και ταυτοποίηση τους. Η δυνατότητα αυτή θα μπορούσε σε επόμενο στάδιο να χρησιμοποιηθεί στον τομέα της τεχνητής νοημοσύνης και της επιχειρηματικής εφυίας για την αναγνώριση αντικειμένων σε τρισδιάστατο χώρο.

Βιβλιογραφία

1. *Big Data as an e-health service.* **Liu, W. and Park, E. K.** s.l. : IEEE, 2014. International Conference on Computing, Networking and Communications (ICNC). pp. 982 - 988.
2. **Dean, Jared.** *Big Data,Data Mining,and Machine learning.* s.l. : wiley, 2014.
3. **McKinsey.** *Big data:The next frontier for innovation,competition,and productivity.* [ed.] Global Institute. 2011.
4. **Juniper networks.** Introduction to Big Data:Infrastructure and Networking Considerations. Σεπτεμβριος 2012, p. 11.
5. **P.Chatterjee.** *Big data:The greater good or invasion of privacy?* 2013. http://www.guardian.co.uk/commentisfree/2013/mar/12/big- data-greater-good-privacy-invasion, 2013..
6. *Big Data Is Opening Doors, but Maybe Too Many.* **S.Lohr.Big Data is Opening doors, But Maybe Too Many.** Μαρτιος 2013, real clear politics, Vol. TECHNOLOGY .
7. *3d data management:Controlling data volume,velocity and variety.* **L.Douglas (Meta group).** 2001, Vol. Application Delivery Strategies, p. 4.
8. **Nist Big Data program.** *NIST Big Data Working Group (NBD-WG). http://bigdatawg.nist.gov/home.php.* 2013.
9. **M.A Beyer and D.Laney.** *The importance of Big data :A definition.Stamford ct.* s.l. : Gartner, 2012.
10. **IBM.** *What is big data? - Bringing big data to the enterprise.* 2013. http://www-01.ibm.com/software/data/bigdata/ .
11. **J.P Dijcks ORACLE.** *Oracle:Big data for the enterprise.* 2013. p. 16.
12. *The Big Bang: How the Big Data Explosion Is Changing the World - Microsoft UK Enterprise Insights Blog - Site Home - MSDN Blogs.* 2013. <http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/0 big-bang-how-the-big-data-explosion-is-changing-the-world.aspx>.
13. **Roberd Hillard.** Big Data Definition - MIKE2.0, the open source methodology for Information Development. *MIKE 2.0.* 2012. http://mike2.openmethodology.org/wiki/Big Data Definition..
14. **Chemawat, Jeffrey Dean and Sanjay.** *MapReduce: simplified data processing on large clusters. In Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation* -. USA : s.n., 2004. Vols. Usenix Association,Berkeley.
15. **Ανδρέας, Παπαδόπουλος.** *Hadoop.* Πληροφορικης, Πανεπιστημιο Κυπρου. Κυπρος : s.n., 2012.
16. *Leaving Data on the Table: Data Scientists Reveal Obstacles to Big Data Analytics.* s.l. : Paradigm4 Data,Scientist Survey. p. 16.
17. **ESRL.** *Geospatial Visualization and Big Data.* . s.l. : Ανάκτηση από ww.esri.com/products/technology-topics/big-data/geospatial-uses, 2011.
18. **Berkeley.** *Big Data Analytics: Making Data Work.* University of California. s.l. : Ανάκτηση από http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html, 2012.

19. **Uma Srinivasan, Bavani Arunasalam.** *Leveraging big data analytics to reduce healthcare costs.* 2013.
20. **Nitesh V. Chawla, Darcy A. Davis.** *Bringing big data to personalized health-care: a patient-centered framework.* s.l. : J. Gen. Intern. Med., 2013. 660–665..
21. **ΡΑΥΤΟΠΟΥΛΟΣ, ΓΕΩΡΓΙΟΣ Ν.** *ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ.* ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ ΤΟΜΕΑΣ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ, ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ . Πατρα : s.n., 2012. p. 96.
22. **Κωνσταντίνος, Γιωτόπουλος.** *Διαφάνειες Μαθήματος.* Διοίκησης Επιχειρήσεων, ΤΕΙ Πάτρας. 2014.
23. **(https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF).**
24. **ΛΑΟΥΡΑ, ΘΕΟΔΟΣΗ-ΚΟΚΚΙΝΟΥ.** *ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΕΦΑΡΜΟΦΕΣ ΣΤΑ ΣΥΣΤΗΜΑΤΑ ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ.* ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ, ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ. ΠΑΤΡΑ : s.n., 2013. ΔΙΠΛΩΜΑΤΙΚΗ.
http://nemertes.lis.upatras.gr/jspui/bitstream/10889/6401/1/%CE%B4%CE%B9%CF%80%CE%BB%CF%89%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%B7.pdf.
25. **57https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF.**
26. **ΓΕΡΑΣΙΜΟΣ, ΑΝΤΖΟΥΛΑΤΟΣ.** *Διαφάνειες Μαθήματος.* ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ , ΤΕΙ ΠΑΤΡΑΣ. 2013.
27. **https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B7%CE%B3%CE%BF%CF%81%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7.**