Επισκόπηση Νέων Ποσοτικών Μεθόδων και Πλεονεκτήματα από την Εφαρμογή τους στον Κόσμο των Επιχειρήσεων

Χρήστος Φράγκος

1 2 3 4 5 6 7 8 9 10 11 12 13

Επισκόπηση Νέων Ποσοτικών Μεθόδων και Πλεονεκτήματα από την Εφαρμογή τους στον Κόσμο των Επιχειρήσεων

Χρήστος Φράγκος

Καθηγητής Τμήματος Διοίκησης Επιχειρήσεων, ΤΕΙ Αθήνας

Abstract

The present paper is a brief introduction to the emerging methodologies in Data Analysis. The economy of 21st century is based on the flow of information and this fact has a profound effect not only on the operation of organizations, but also on the environment of business analysis.

The tremendous amount of information generated by today's companies, private or state-owned, will only increase. As electronic commerce and electronic communication have become more common on the INTERNET, firms are developing a direct link with customers and this is another source of information. There are two important areas which are the tasks of every research worker Data Management and Analysis of Data. **The techniques of DATA WAREHOUSING and DATA MINING** greatly help the task of Data Management.

During the last twenty hears the important family of **SUB-SAMPLING STATISTICAL TECHNIQUES** has been developed, due to the energetic efforts of B. Efron and a group of other eminent research workers, [J. SWANEPOEL, C.SWANEPOEL,M. KNOTT,among others]. These techniques, which include **the BOOTSTRAP** and **JACKKNIFE COMPUTER INTENSIVE METHODS**, greatly facilitate the Analysis of Data today. The present paper is an introduction to the above techniques and aims to identify their advantages in the analysis of business data.

Το άρθρο αυτό έχει σκοπό να παρουσιάσει τις αναδυόμενες μεθοδολογίες της τελευταίας εικοσαετίας στον τομέα των Ποσοτικών Μεθόδων.

Οι τεχνικές αυτές είναι: Τενικές Επαναχρησιμοποίησης του δείγματος [resampling techniques, bootstrap και jackknife techniques], εξόρυξη δεδομένων [data mining and warehousing] και Νευρωνικά Δίκτυα [neural networks].

Η βάση της Οικονομίας του 21ου αιώνα είναι η πληροφορία. Το γεγονός αυτό έχει σημαντικές επιπτώσεις,όχι μόνο στις συναλλαγές των δημοσίων και ιδιωτικών βιομηχανιών και επιχειρήσεων αλλά και στο περιβάλλον για εμπορική ανάλυση. Οι σημερινοί ερευνητές των προβλημάτων των επιχειρήσεων αλλά και εκείνοι που ερευνούν θεωρητικά προβλήματα αντιμετωπίζουν νέες προκλήσεις που απαιτούν την χρήση παραδοσιακών Στατιστικών Τεχνικών σε συνδυασμό με την εφαρμογή νέων αναλυτικών μεθόδων.

Το τεράστιο μέγεθος των πληροφοριών που παράγονται από τις σημερινές επιχειρήσεις, θα αυξάνεται με εκθετική τάση. Καθώς το ηλεκτρονικό εμπόριο και η ηλεκτρονική επικοινωνία δια μέσου του Internet έχουν γίνει κοινή πρακτική, οι επιχειρήσεις αναπτύσσουν διαύλους άμεσης επικοινωνίας με τους πελάτες τους, οι οποίοι είναι μια άλλη πηγή πληροφοριών.

Ο σημερινός ερευνητής έχει δύο σοβαρές εργασίες να διεκπεραιώσει. Διαχείριση Δεδομένων [Data Management] και Ανάλυση Δεδομένων [Analysis of Data]. Η διαχείριση δεδομένων χρησιμοποιεί την τεχνική της αφομοίωσης των δεδομένων [Data Warehousing, Data Assimilation], η οποία είναι τεχνική ενσωμάτωσης των δεδομένων [Integration of Data] σε πληροφοριακή ροή η οποία είναι κατάλληλη για αναλυτική επεξεργασία. Πολύ συγγενής προς την τεχνική της ενσωμάτωσης των δεδομένων είναι η εξόρυξη δεδομένων [Data Mining].

Η Ανάλυση των δεδομένων έχει γίνει πιο διερευνητική με τη χρήση των μοντέρνων μεθοδολογιών της επαναχρησιμοποίησης του δείγματος. Αυτές είναι οι τεχνικές Bootstrap και Jackkuife που περιλαμβάνονται στην οικογένεια των τεχνικών επαναχρησιμοποίησης του δείγματος [Resampling Techniques].Τέλος, η ανάλυση των Δεδομένων έχει ενεργετηθεί πολύ από την χρήση των μοντέλων εκμάθησης

[learning models],ένα παράδειγμα των οποίων είναι τα Νευρωνικά Δίκτυα [Neural Networks].

Η παρούσα εργασία είναι μια περιληπτική εισαγωγή στις παραπάνω μεθοδολογίες με έμφαση στις τεχνικές επαναχρησιμοποίησης τού δείγματος.

Λέξεις- Κλειδιά: Data Mining, Data Warehousing, Resampling techniques, Bootstrap method, Jackknife method, Neural Networks.

1. Εισαγωγή

Τα θεμελιώδη, χαρακτηριστικά της Ανάλυσης Δεδομένων παρέμειναν στατικά μέχρι τα μέσα της δεκαετίας του 1970.

Το 1976 αρχίζει η συστηματική εισαγωγή των Η/Υ μεγάλης δυναμικότητας των εταιρειών IBM, COC, HONEYWELL BULL, CRAY, ICL και άλλων, στα Πανεπιστήμια, στα Ερευνητικά Ιδρύματα και στις επιχειρήσεις. Παράλληλα, λόγω κόστους και μεγαλύτερης ευκολίας χρήσης,αρχίζει η εισαγωγή και η εξέλιξη των προσωπικών υπολογιστών. Οι ερευνητές λοιπόν έχουν, από το σημείο αυτό, αρκετή υπολογιστική δυνατότητα εκατομμυρίων υπολογισμών το δευτερόλεπτο η οποία,ουσιαστικά,εξαφάνισε όλους τους περιορισμούς στους τύπος των στατιστικών τεχνικών που είναι στη διάθεση τους.

Η προσοχή των ερευνητών τώρα στρέφεται όχι μόνο στη θεωρητική υποδομή των μονοδιάστατων και πολυδιάστατων Ποσοτικών Τεχνικών αλλά και στην φύση και του χαρακτήρα των δεδομένων.

Η δεύτερη Επανάσταση στην Ανάλυση Δεδομένων είναι η εισαγωγή της «εποχής της Πληροφορίας» το 1985-1990.

Μεγάλες βάσεις δεδομένων με εκατοντάδες χιλιάδων ή και εκατομμύρια δεδομένων δίνουν πολύ λεπτομερείς εικόνες των χαρακτηριστικών ενός πληθυσμού για πολλά χρόνια. Υπάρχει έκρηξη δεδομένων η οποία αλλάζει πλήρως την γωνία της εφαρμογής των ποσοτικών τεχνικών στο ερευνητικό περιβάλλον.

Το ερώτημα τίθεται: Ποιες ποσοτικές τεχνικές είναι οι πιο κατάλληλες για την νέα πρόκληση της εποχής της πληροφορίας; Πως οι πληροφορίες θα τύχουν επεξεργασίας και θα αναλυθούν έγκαιρα και συστηματικά;

Η απάντηση των ερευνητών είναι η τάση να γίνεται ανάλυση χωρίς στατιστική συμπερασματολογία. **Υπάρχει η ροπή να 'γυρίσουμε στα δεδομένα'** και να εφαρμοσθούν από τον ερευνητή όσο το δυνατόν λιγότεροι περιορισμοί για την ανάλυση. **Κατόπιν 'ας αφήσουμε τα δεδομένα να μιλήσουν'**

«let's the data talk». Στο κλίμα αυτό άεχουν αναπτυχθεί, αφ'ενός μεν τα προγράμματα τής τεχνητής νοημοσύνης των οποίων είναι παραπροϊόντα τα Νευρωνικά Δίκτυα και οι Γενετικοί Αλγόριθμοι και αφ'ετέρου οι μεθοδολογίες επαναχρησιμοποίησης του δείγματος των οποίων αντιπροσωπευτικά μέλη είναι οι **Bootstrap και Jackknife Στατιστικές Μέθοδοι σημειακής εκτίμησης παραμέτρων και κατασκευής διαστημάτων εμπιστοσύνης.**

2. Βασικοί Ορισμοί

Δίνουμε τους παρακάτω ορισμούς:

a. **Αφομοίωση Δεδομένων (Data warehousing)** Είναι η προσπάθεια να συνδυασθούν όλες οι πηγές των δεδομένων και οι υπάρχουσες πληροφορίες που είναι σχετικές σε έναν οργανισμό σε μία μοναδική, ενοποιημένη βάση δεδομένων με δομή κατάλληλη για την υποστήριξη ανάλυσης και λήψης αποφάσεων από όλα τα επίπεδα του οργανισμού.

b. **Εξόρυξη Δεδομένων (Data Mining)** Είναι η ερευνητική περιοχή της Ανάλυσης Δεδομένων που περιλαμβάνει όλες τις τεχνικές που εξερευνούν τα δεδομένα και ανακαλύπτουν κοινά χαρακτηριστικά και κοινές σχέσεις.

Οι Στατιστικές και Ποσοτικές Τεχνικές που χρησιμοποιούνται στην εξόρυξη Δεδομένων είναι η Γραφική Ανάλυση Πολυδιάστατων Δεδομένων, η Ανάλυση Συστάδων (cluster Analysis), η Παραγοντική Ανάλυση (factor Analysis), η Πολλαπλή και η Λογιστική Παλινδρόμηση (Multiple or Logistic Regression).

c. **Επαναχρησιμοποίηση Δείγματος (Resampling Technique) Είναι η τεχνική της εμπειρικής εκτίμησης των δειγματικών Κατανομών των στατιστικών μεγεθών μέσω της διαδοχικής τυχαίας δειγματοληψίας με επανατοποθέτηση από το αρχικό δείγμα.** Η υπολογιστική, δυνατότητα των σημερινών συστημάτων Η/Υ που μπορούν να εκλέγουν χιλιάδες τυχαία δείγματα με επανατοποθέτηση από το αρχικό δείγμα σε κλάσματα λεπτού καθιστά ικανό τον ερευνητή να αποφύγει της υποθέσεις της Κανονικής Κατανομής των δεδομένων. Με αυτό τον τρόπο, ο ερευνητής, **πραγματικά 'ξαναγυρίζει στα δεδομένα'** και εκτιμά τα πραγματικά χαρακτηριστικά των δεδομένων χωρίς να υποθέτει ότι ισχύουν σουρισμένες κατανομές για το σκοπό της εκτίμησης.

Οι τεχνικές της Αφομοίωσης των Δεδομένων και της εξόρυξης Δεδομένων παρουσιάζονται στις εργασίες των Banquin, R. και Edelstein, H. (1996), Simon, A.R. (1995), Boar, B. (1996) και Berry, M., και Linoff, G. (1997).

Η τεχνική των Νευρωνικών δικτύων παρουσιάζεται στις εργασίες των Bigus, J. (1996), and Chester, M. (1993), Fausett, L. (1994), Hertz, J. (1991) and Smith, M. (1993).

Η οικογένεια των τεχνικών επαναχρησιμοποίησης του δείγματος παρουσιάζεται στις εργασίες των Efron, B. (1982), Efron, B. And Tibshirani, R. J. (1993), Frangos, C.C. (1986, 1995) και Mooney, C.Z.και Duval, R.D. (1993).Τα επόμενα κεφάλαια δίνουν τις βασικές ιδέες των παραπάνω τεχνικών, την ορολογία, με μερικά παραδείγματα στον κόσμο των επιχειρήσεων.

3. Αφομοίωση Δεδομένων και Εξόρυξη Δεδομένων

Ο σκοπός της Αφομοίωσης και εξόρυξης Δεδομένων είναι η βελτίωση της πρόσβασης δεδομένων για λήψη αποφάσεων. Η αφομοίωση δεδοδένων είναι ο μηχανισμός που διευκολύνει τα συστήματα Υποστήριξης αποφάσεων (Decision Support Systems, DSS) αποθηκεύοντας τα δεδομένα και προμηθεύοντας μια ιστορική προοπτική των δεδομένων. **Δύο βασικές έννοιες στην αφομοίωση δεδομένων είναι η ενοποίηση (integration) και η χρονική έλλειψη μεταβλητικότητας (time invariance).**

Ενοποίηση είναι η αποθήκευση όλων των δεδομένων σε μια ενοποιημένη βάση δεδομένων η οποία συγκεντρώνει όλες τις πηγές των δεδομένων μιας επιχείρησης σε ένα μοναδικό σημείο πρόσβασης.

Χρονική έλλειψη μεταβλητικότητας είναι η ιδιότητα τών αφομοιωμένων δεδομένων να αποτελούν '**φέτες της πραγματικότητας**' που είναι διαθέσιμα σε κάθε αναδρομική ανάλυση.

Η Εξόρυξη Δεδομένων λέγεται επίσης: ανακάλυψη γνώσης σε μεγάλες βάσης δεδομένων (**knowledge discovery in databases, KDD**),είναι η έρευνα για σχέσεις και ομάδες δεδομένων με κοινά χαρακτηριστικά σε μεγάλες βάσεις δεδομένων.

Μερικά παραδείγματα εξόρυξης Δεδομένων είναι τα ακόλουθα:

3α Παραδείγματα Εφαρμογών Εξόρυξης Δεδομένων, Νευρωνικών Δικτύων και Μεθόδου Bootstrap.

Α. Μια Αμερικανική Εταιρεία, η Camelot Music Holdings χρησιμοποίησε εξόρυξη δεδομένων για να επισημάνει μια ομάδα αγοραστών μεγάλης αγοραστικής συχνότητας οι οποίοι ήταν ηλικίας 65 χρόνων και άνω. Οι αγοραστές αυτοί αγόραζαν αρκετούς cd's με κλασσική και *τζάζ* μουσική αλλά και DVD's με έργα κινηματογραφικά. Μια πιο επίμονη εξόρυξη δεδομένων απεκάλυψε ότι ένα μεγάλο ποσοστό των αγοραστών ηλικίας 65 ετών και άνω

αγόραζαν μουσική **rap** και εναλλακτικές μορφές. Τι συνέβαινε: Οι αγοραστές αυτοί ήταν γιαγιάδες και παππούδες που αγόραζαν για τα εγγόνια τους. Έκτοτε, οι πωλήσεις της Camelot, αυξήθηκαν με την εισαγωγή διαφημιστικών φυλλαδίων στους πελάτες της τρίτης ηλικίας με τα νέα κομμάτια στην μουσική **rap** και τις εναλλακτικές μορφές μουσικής.

Β. Η τράπεζα American Express χρησιμοποιεί Νευρωνικά δίκτυα για να εξετάσει τις εκατοντάδες των εκατομμυρίων εγγραφές στη βάση δεδομένων που διατηρεί για να ανακαλύψει πώς και πού οι κάτοχοι κάρτας κάνουν συναλλαγές. Το Νευρωνικό δίκτυο είναι ένα πρόγραμμα που μιμείται τις διεργασίες του ανθρωπίνου μυαλού και συνεπώς είναι ικανό να μάθει από παραδείγματα και να βρίσκει ομάδες δεδομένων με κοινά χαρακτηριστικά. Το αποτέλεσμα της ανίχνευσης ήταν η καθιέρωση συνόλου βαθμών στην αγοραστική τάση για κάθε κάτοχο κάρτας. Βασισμένη στους βαθμούς αυτούς η American Express συνδυάζει διαφορές στη αγοραστική ιστορία κάθε κατόχου κάρτας και περιλαμβάνει τις προσφορές αυτές με τους μηνιαίους λογαριασμούς. Το αποτέλεσμα είναι η μεγαλύτερη χρήση των καρτών American Express και πληροφορίες ειδικών προσφορών καταστημάτων που ενδιαφέρουν τους κατόχους κάρτας.

Γ. Ένα παράδειγμα εφαρμογής της μεθόδου Bootstrap είναι η κατασκευή διαστημάτων εμπιστοσύνης για τον συντελεστή συσχέτισης χωρίς να υπάρχουν ιδιαίτερες υποθέσεις για το είδος της κατανομής τον συντελεστή συσχέτισης.

4. Διαχείριση μίας Αποθήκης Δεδομένων

Η διαχείριση μίας αποθήκης δεδομένων περιλαμβάνει τα εξής στάδια:

1. **Λήψη δεδομένων** από όλες τις εσωτερικές ή εξωτερικές πηγές μιας επιχείρησης (π.χ. λογιστήριο, αποθήκη, δημογραφικά δεδομένα πελατών, έρευνες αγοράς κ.τ.λ.)
2. **Ενοποίηση δεδομένων**. Στο στάδιο αυτό τα δεδομένα πού έρχονται από διάφορες πηγές ενοποιούνται, διατηρώντας την συνέπεία τους με το ταίριασμα των χαρακτηριστικών και των ιδιοτήτων.
3. **Καθαρισμός Δεδομένων (data cleaning)**. Στο στάδιο αυτό εξετάζονται τα δεδομένα για την ύπαρξη σφαλμάτων και γίνονται έλεγχοι συνεπείας.
5. **Δημιουργία μεταδεδομένων (metadata creation)**.
   Η φάση αυτή περιλαμβάνει την δημιουργία ενός στοιχείου δεδομένων που περιλαμβάνει την αρχική πηγή και όλους τους μετασχηματισμούς των δεδομένων.

6. **Εισαγωγή δεδομένων.** Περιοδική εισαγωγή δεδομένων στη βάση δεδομένων και ενοποίηση τους με τα υπάρχοντα δεδομένα.
7. **Τακτοποίηση δεδομένων.** Οργάνωση δεδομένων.
8. **Υποστήριξη αποφάσεων.** Το στάδιο αυτό αναφέρεται στην λήψη αποφάσεων με τη χρήση της βάσης δεδομένων χρησιμοποιώντας το **σύστημα OLAP (on- line analytical processing)** ή σε διαδικασίες εξόρυξης γνώσης από την υπάρχουσα βάση δεδομένων.

4α Συστήματα Οργανισμών.

Υπάρχουν δύο ειδών συστήματα σε κάθε επιχείρηση

1) Επιχειρησιακά Συστήματα (operational Systems) (OS) Αυτά είναι τα συστήματα ελέγχου του λογιστηρίου, των αποθηκών, του τμήματος παραγγελιών.

2) Συστήματα υποστήριξης αποφάσεων (DSS)

Συνεπώς υπάρχουν δύο ειδών βάσεις δεδομένων,

1) **Πρόσφατα δεδομένα**
2) **Ιστορικά Δεδομένα**

Στην διαδικασία της ενοποίησης δεδομένων συναντώνται οι ακόλουθοι τύποι δεδομένων

- **επιχειρησιακά δεδομένα (operational data)** σε αντίθεση με τα αναλυτικά δεδομένα
- **Πρωτογενή δεδομένα (primitive data)** σε αντίθεση με τα συσσωρευτικά δεδομένα (aggregated data)
- **μεταδεδομένα (metadata)** τα οποία έχουν τυποποιημένο σύστημα ταξινόμησης μέσα στην επιχείρηση.

Μετά την ενοποίηση των δεδομένων (data warehousing) έρχεται η διαδικασία της **εξόρυξης γνώσης (data mining)** η οποία είναι μία άλλη προσέγγιση στην ανάλυση δεδομένων. **Είναι μια άλλη διαδικασία που χρησιμοποιεί πολλές πολυδιάστατες ποσοτικές τεχνικές, όπως cluster analysis, factor analysis, MANOVA, για να ανακαλύψει ομάδες δεδομένων με κοινά χαρακτηριστικά. Επίσης χρησιμοποιούνται οι τεχνικές multiple regression, Discriminant Analysis and Logistic Regression**

5. Computer Programs for Data Mining

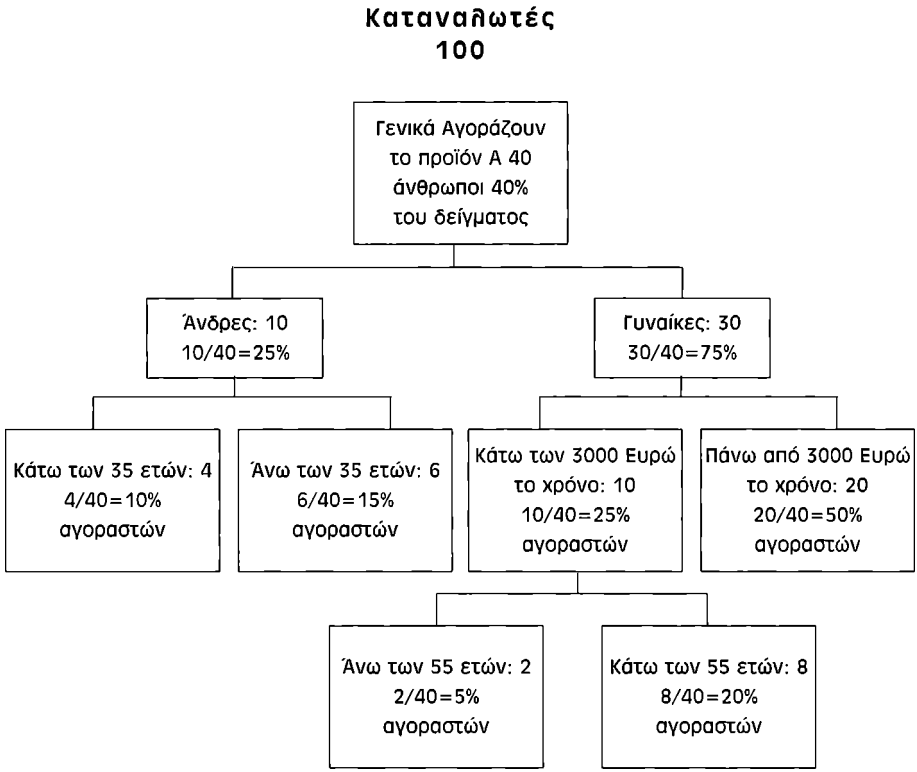
1. **The programs developed by SPSS: CLEMENTINE**
2. **The programs developed by IBM: DIAMOND**

3. The programs developed by SAS: data mining
4. decision trees programs : CHAID
5. decision trees programs : CART

5α. Παράδειγμα δένδρου αποφάσεων

Ένα δείγμα καταναλωτών περιλαμβάνει 100 καταναλωτές άνδρες και γυναίκες.

Σε κάθε κουτί αναφέρεται η ονομασία του τμήματος των αγοραστών και το ποσοστό των αγοραστών.



Άρα, το μεγαλύτερο ποσοστό των αγοραστών είναι γυναίκες με εισόδημα πάνω από 30000 ¤ ετήσια. Αποφασίζει λοιπόν η εταιρία να δώσει σε κάθε γυναίκα πελάτισσα με εισόδημα πάνω από 30000 ¤ το χρόνο μια ειδική προσφορά για να αυξήσει τις πωλήσεις του προϊόντος Α.

6. Νευρωνικά Δίκτυα (NEURAL NETWORKS)

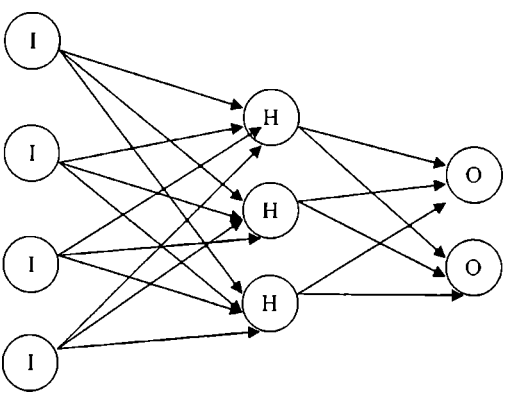
Τα Νευρωνικά δίκτυα είναι ένα από τα εργαλεία τα οποία είναι πιο πιθανό να συσχετισθούν με την εξόρυξη γνώσης. Τα Νευρωνικά δίκτυα είναι βασισμένα επάνω στην διαδικασία εκμάθησης του ανθρώπινου μυαλού και προσπαθεί να "μάθει" με επαναλαμβανόμενες δοκιμές τον τρόπο με τον οποίο θα οργανωθεί με τον καλύτερο τρόπο για να επιτύχει την καλύτερη πρόβλεψη.

Η Βασική λειτουργία ενός νευρωνικού δικτύου είναι η εξής:

Το μοντέλο αποτελείται από **σταθμούς επεξεργασίας (nodes)** οι οποίοι είναι εισροές, εκροές ή ενδιάμεσοι επεξεργαστές. Κάθε επεξεργαστής συνδέεται με το επόμενο σύνολο επεξεργαστών με μία σειρά **σταθμικών βημάτων (weighted path-weights)** (Αυτά τα σταθμικά βήματα είναι παρόμοια με τους συντελεστές βαρύτητας σε ένα σταθμισμένο μοντέλο παλινδρόμησης). Βασισμένο στην διαδικασία μάθησης, το μοντέλο λαμβάνει την πρώτη περίπτωση (το πρώτο δείγμα), θέτει τα δεδομένα στην **συνάρτηση ενεργοποίησης (activation function)** και λαμβάνει μια αρχική απόφαση, βασισμένο στους συντελεστές βαρύτητας. Υπολογίζεται το σφάλμα πρόβλεψης και μετά το μοντέλο κάνει την καλύτερη προσπάθεια να αλλάξει τους συντελεστές βαρύτητας για να βελτιώσει την πρόβλεψη και μετά λαμβάνει την δεύτερη περίπτωση (δεύτερο δείγμα).

Ο κύκλος αυτός επαναλαμβάνεται για κάθε περίπτωση στην φάση δοκιμής (training phase) όταν το μοντέλο προσαρμόζεται. Μετά την προσαρμογή το μοντέλο μπορεί να χρησιμοποιηθεί σε ένα ξεχωριστό δείγμα για να αξιολογηθεί.

Ένα νευρωνικό δίκτυο έχει την ακόλουθη μορφή:

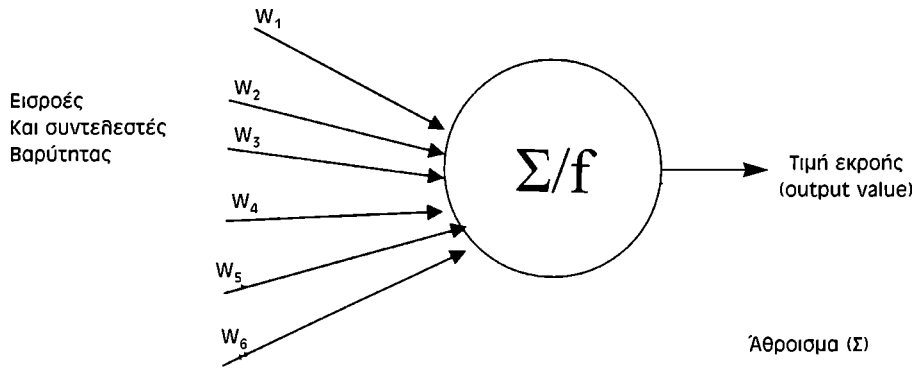


Σταθμοί Εκροών (Output nodes) (O)

Σταθμοί Εισροών (Input nodes) (I)

Κρυφοί (Hidden nodes) (H)

Σταθμός (Node)



Τα Νευρωνικά δίκτυα μπορούν να αντιμετωπίζουν προβλήματα πρόβλεψης και ταξινόμησης. Για λεπτομερέστερη περιγραφή των Νευρωνικών δικτύων, ο αναγνώστης παραπέμπεται στις εργασίες των Bigus, J. (1996), Chester, M.(1993) και Fausett, L(1994)

7. Τεχνικές Αναδειγματοληψίας (Resampling Techniques)

Οι τεχνικές επαναδειγματοληψίας αγνοούν την υποτιθέμενη **δειγματική Κατανομή (sampling distribution**) ενός στατιστικού (statistic) το οποίο είναι εκτιμητής μίας παραμέτρου (**estimate of a parameter**) και υπολογίζουν μια εμπειρική κατανομή, η οποία είναι η πραγματική κατανομή τού στατιστικού, χρησιμοποιώντας εκατοντάδες ή και χιλιάδες δείγματα. Από κάθε ένα από τα δείγματα αυτά υπολογίζεται το στατιστικό και το ενενηκοστό πέμπτο ή ενενηκοστό ένατο εκατοστιαίο σημείο του στατιστικού. Έτσι υπολογίζεται το διάστημα εμπιστοσύνης για την παράμετρο που εκτιμά το στατιστικό με συντελεστή 95% ή 99% αντίστοιχα.

Από πού προέρχονται τα εκατοντάδες δείγματα;

Τα εκατοντάδες δείγματα προέρχονται από το αρχικό δείγμα με αναδειγματοληψία. Στην περίπτωση της τεχνικής που λέγεται **bootstrap** η αναδειγματοληψια είναι τυχαία με επανάθεση και στην περίπτωση της τεχνικής που λέγεται **Jackknife,** η αναδειγματοληψία είναι συστηματική,με επανάθεση, εξαιρώντας κάθε φορά μία παρατήρηση από το αρχικό δείγμα μεγέθους (n) και λαμβάνοντας δείγμα μεγέθους (n-1).

7α Παραδείγματα των τεχνικών Bootstrap και Jackknife σε προβλήματα εκτιμητικής.

Υποθέτουμε ότι τα δεδομένα είναι $Z_1, Z_{2,}....,Z_{15}$ όπου $Z_i = (X_i, \Psi_i)$, i=1,2,...,15

Οι X και Ψ είναι οι συσχετισμένες μεταβλητές όπως

X_i= Πωλήσεις ενός Supermarket στον χρόνο i

Ψ_i= Ποσό που διέθεσε ο διευθυντής του Supermarket για διαφημίσεις στον χρόνο i, i=1,2,...,15

Ζητείται να εκτιμηθεί ο συντελεστής συσχέτισης ρ_{XY} και να υπολογισθεί ο εκτιμητής της τυπικής απόκλισης του

Ακολουθούνται τα εξής βήματα

1. Χρησιμοποιώντας μια γεννήτρια τυχαίων αριθμών, παράγονται οι τυχαίοι αριθμοί. $i_1, i_2,...., i_n$. Θέτουμε $Z^*_1 = Z_{(i1)} = (X_{(i1)}, \Psi_{(i1)})$

$$Z^*_2 = Z_{(i2)} = (X_{(i2)}, \Psi_{(i2)})$$

$$Z^*_n = Z_{(in)} = (X_{(in)}, \Psi_{(in)})$$

Από το δείγμα $Z^*_1, Z^*_2,...., Z^*_n$, υπολογίζεται ο συντελεστής συσχέτισης του PEARSON χρησιμοποιώντας τον τύπο:

$r_{XY,b}$ =

$$\frac{\sum_{\kappa=1}^{n} X_\kappa Y_\kappa - (\sum_{\kappa=1}^{n} X_\kappa)(\sum_{\kappa=1}^{n} Y_\kappa)}{(\sum_{\kappa=1}^{n} X_\kappa^2 - (\sum X_k)^2)^{1/2} (\sum_{\kappa=1}^{n} Y_\kappa^2 - (\sum Y_k)^2)^{1/2}}$$

2. Επαναλαμβάνεται το βήμα 1, B φορές (B=100 έως 1000) Υπολογίζονται οι τιμές του συντελεστή $r_{XY,b}$ από κάθε bootstrap δείγμα,b=1,...,B
3. Ο εκτιμητής bootstrap του συντελεστή συσχέτισης ρ_{XY} είναι

$$\rho_{XY,B} = \frac{\sum_{b=1}^{B} r_{XY,b}}{B}$$

με εκτιμητή τυπικής απόκλισης

$$\hat{\sigma}_{\rho,B} = \left(\frac{\sum_{b=1}^{B}\{r_{XY,b} - \rho_{XY,B}\}^2}{B-1}\right)^{\frac{1}{2}}$$

Είναι εύκολο να αποδειχθεί (Efron, 1982,1986) ότι καθώς το B→∞, το $\hat{\sigma}_{\rho,B}$ θα πλησιάσει σ_ρ, τη τυπική απόκλιση του ρ.

Συνεπώς, η μέθοδος bootstrap, έλυσε, με την δύναμη του υπολογιστή,το αναλυτικό πρόβλημα της εύρεσης της αριθμητικής τιμής της τυπικής απόκλισης του δειγματικού συντελεστή συσχέτισης ρ_{XY} του Pearson.

7b. Παράδειγμα εφαρμογής της μεθόδου Jackknife

Όλοι γνωρίζουμε ότι η κατασκευή ενός διαστήματος εμπιστοσύνης για την διακύμανση σ^2 προυποθέτει τη συνθήκη: οι παρατηρήσεις $x_1, x_2, \ldots x_n$ να ακολουθούν την Κανονική Κατανομή.

Έστω τώρα ότι ένα δείγμα $y_1, y_2, \ldots, y_n$ παρατηρήσεων δεν προέρχεται απαραίτητα από την Κανονική Κατανομή. Υποθέτουμε ότι ο είναι ένας εκτιμητής της παραμέτρου θ που είναι συνάρτηση του δείγματος μεγέθους n. Έστω $\hat{\theta}_1$ ο αντίστοιχος εκτιμητής της θ που είναι συνάρτηση του δείγματος $y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ μεγέθους n-1, (i=1,2,.....,n)

Ορίζουμε:

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$$

(i=1,2,.....,n)

Ο εκτιμητής Jackknife της παραμέτρου θ είναι:

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\hat{\theta}_{-i}}{n}$$

και έχει της εξής δύο σπουδαίες ιδιότητες:

1. Μειώνει την μεροληψία του εκτιμητή κατά το κλάσμα 1/n

2. Ο μη παραμετρικός εκτιμητής της διακύμανσης του αρχικού εκτιμητού ,όπως και του $\hat{\theta}$ εκτιμητού Jackknife $\hat{\theta}$ είναι:

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})^2$$

3. Το κλάσμα

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i-\hat{\theta})^2\right\}^{\frac{1}{2}}}$$

ακολουθεί, προσέγγιστικά την (t) κατανομή με n-1 βαθμούς ελευθερίας (Miller, 1974)

Άρα ,μπορεί να κατασκευασθεί το διάστημα εμπιστοσύνης για την παράμετρο θ,χωρίς να απαιτείται γνώση της Κατανομής των αρχικών παρατηρήσεων y1,y2,...,yn

7c. Παράδειγμα εφαρμογής των στατιστικών μεθόδων Jackknife και Bootstrap για τη κατασκευή διαστημάτων εμπιστοσύνης για τη παράμετρο κήσης του Λογιστικού Μοντέλου

Θεωρούμε το Λογιστικό Μοντέλο:

$$\Theta_j=\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

όπου θj =P(Yij=1)

1-θj=P(Yij=0) (i=1,2,...,k) (j=1,2,...,n)

Οι μέθοδοι Bootstrap και Jackknife εφαρμόζονται για να κατασκευασθούν 95% Διαστήματα

Εμπιστοσύνης για τη παράμετρο β1. Ο παρακάτω πίνακας Α περιέχει την πιθανότητα κάλυψης

(coverage probability),το μέσο μήκος διαστήματος(expected length) και τη τυπική απόκλιση

του μέσου μήκους των Διαστημάτων Εμπιστοσύνης για τη παράμετρο β1. Τα Διαστήματα αυτά παράγονται με τη βοήθεια της μεθόδου Jackknife (F1) και τριών παραλλαγών της μεθόδου

Bootstrap(J1,J2,J3)

Επισημαίνεται ότι η μέθοδος J3 επιτυγχάνει την κατασκευή Διαστημάτων Εμπιστοσύνης για μέγεθος δείγματος n=5,n=10 και αριθμό δειγμάτων Bootstrap k=20,30,40,με τα εξής χαρακτηριστικά

Α. Πολύ καλή πιθανότητα κάλυψης.

Β. Ελάχιστο μέσο μήκος Διαστήματος.

Γ. Μικρή τυπική απόκλιση.

Πίνακας Α: Πιθανότητα κάλυψης(C),Μέσο Μήκος(L) και τυπική απόκλιση Μέσου των Διαστημάτων Εμπιστοσύνης για τη παράμετρο β1 του Λογιστικού Μοντέλου

		n=5			n=10		
k		C	L	υπ.απ.	C	L	τυπ.απ.
20	F1	0,94	0,756	(0,017)	0,94	0,534	(0,003)
	J1	0,90	0,767	(0,027)	0,93	0,536	(0,004)
	J2	0,90	0,760	(0,021)	0,94	0,532	(0,003)
	J3	0,93	0,684	(0,019)	0,94	0,460	(0,003)
	F1	0,94	0,612	(0,005)	0,95	0,424	(0,002)
	J1	0,93	0,610	(0,006)	0,96	0,425	(0,003)
	J2	0,93	0,612	(0,006)	0,95	0,428	(0,003)
	J3	0,94	0,556	(0,004)	0,95	0,374	(0,002)
	F1	0,94	0,521	(0,004)	0,94	0,372	(0,002)
	J1	0,93	0,524	(0,004)	0,94	0,368	(0,002)
	J2	0,93	0,522	(0,004)	0,94	0,370	(0,002)
	J3	0,94	0,475	(0,004)	0,95	0,338	(0,002)

Συμπέρασμα

Είμαστε σε μία εποχή έντονης αλλαγής στις Ποσοτικές μεθόδους. Η ανακάλυψη και εξάπλωση των Η/Υ έφερε την εποχή της πληροφορίας και τις τεχνικές της εξόρυξης Δεδομένων, των Νευρωνικών Δικτύων και τις Αναδειγματικές τεχνικές Bootstrap και Jackknife.Πολλές από τις τεχνικές αυτές αναδύθηκαν από ανάγκες των εφαρμοσμένων πεδίων της Πολυδιάστατης Ανάλυσης . Ας ελπίσουμε ότι, στο μέλλον, η ανακάλυψη των τεχνικών αυτών θα σημάνει την μεγαλύτερη συνεργασία μεταξύ θεωρητικών και εφαρμοσμένων επιστημόνων των Ποσοτικών Μεθόδων για την πρόοδο της επιστήμης

Βιβλιογραφία

Balakrishanan, P.V., M. C. Cooper, V.S. Jacob, and P.A.Lewis (1994), "A Study of the Classification Capabilities of Neural Networks Using Unsupervised Learning: A Comparison with K-means Clustering " Psyhomertika 59: 509-25

Banquin, R., and H. Edelstein, Eds. (1996), Planning and Designing the Data Warehouse. Upper Saddle River, N.J.: Prentice Hall.

Berry, M., and G. Linoff (1997), Data Mining Techniques for Marketing, Sales and Customer Support. New York: Wiley.

Bigus, J. (1996), Data Mining With Neural Networks. New York: McGraw-Hill.

Bigus, J. (1996), Data Mining With Neural Networks: Solving Business Problems – from Application Development to Decision Support. New York: McGraw-Hill.

Boar, Bernard (1996), Understanding Data Warehousing Strategically. Lincroft, N.J.: NCR Corporation

Chester, M. (1993) Neural Networks: A Tutorial. Upper Saddle River, N.J.: Prentice Hall

Devlin, B. (1977), Data Warehouse: From Architecture to Implementation. New York. Addison-Wesley.

Efron, B. (1982), The Jackknife, the Bootstrap and Other Resampling Plans, vol. 38 of the CBSM-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., and R.J. Tibshirani(1993) An Introduction to the Bootstrap. New York: Chapman and Hall.

Fausett, L. (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Upper Saddle River, N.J.: Prentice Hall.

Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (1996), Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: AAAI Press/ MIT Press.

Frangos,C.C.(1980),"Variance Estimation for the SecondOrdeJackknife, "Biometrika, 67, 3, 715-718.

Frangos, C.C. and M. Knott. (1983), "Variance Estimation for the Jackknife using von Mises Expansions," Biometrika, 70, 2, 501-504.

Frangos, C.C.and M. Stone, (1984), "On Jackknife, Cross-Validatory and Classical Methods of Estimating a Proportion with Batches of Different Sizes,"Biometrika, 71, 2, 361-366.

Frangos, C.C. (1987), "An Updated Bibliography on the Jackknife Method," Communications in Statistics, Theory and Methods, 16, 6, 1543-1584.

Frangos, C.C. and W.R. Schucany (1990), "Jackknife Estimation of the Bootstrap Acceleration Constant," Computational Statistics and Data Analysis,9,1-11.

Frangos, C.C. and C. Swanepoel (1994), "Bootstrap Confidence Limits for the Slope Parameter of the Logistic Model," Communications in Statistics, Simulation and Computation, 23, 4,1115-1126

Frangos, C.C. and W.R. Schucany. (1995), "Improved Bootstrap Confidence Intervals in Certain Toxicological Experiments," Communications in Statistics,Theory and Methods,24,3,829-844.

Groth, R. (1997) Data Mining: A Hands-On Approach for Business Professionals. Upper Saddle River, N.J.: Prentice Hall

Hackathorn, D. (1995) "Reinventing Enterprise Systems via Data Warehousing" The Data Warehousing Institute Annual Conference, Washington, D.C.

Hertz, J. (1991) Introduction to the Theory of Neural Computing. Reading, Mass.: Addison-Wesley

Hinton, G. E. (1992), "How Neural Networks Learn from Experience" Scientific American 267 (September): 144-51.

Holsheimer, M., and M. Kersten (1994), Architectural support for Data Mining. Technical Report CS-R9429. Amsterdam: CWI.

Inmon, W. H., and (1996) Building the Data Warehouse. London: QED Publishing Group.

Inmon, W.H., and R.D. Hackathorn (1994), Using the Data Warehouse. New York: Wiley

Inmon, W.H., J.D Welch, and K. Glassey (1987) Managing the Data Warehouse. New York: Wiley

Kimball, R (1996) The Data Warehouse Toolkit. New York: Wiley

Medsker, L., and J. Liebowitz (1994). Design and Development of Expert Systems and Neural Networks. New York: Macmillan

Michalewicz, Z. (1994) Genetic Algorithms+ Data Structures= Evolution Programs. New York: Springer-Verlag.

Michie, D., D. J. Spiegelhalter, and C.C. Taylor (1994) Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood.

Mooney, C. Z., and R. D. Duval (1993), Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, Calif.: Sage.

Piatetsky-Shapiro, G., and W. Frawley (1991) Knowledge Discovery in Databases. Cambridge, Mass.: AAAI Press/MIT Press.

Redman, T.C (1996), Data Quality for the information Age. New York: Artech House

Simon, A.R. (1995), Strategic Database Technology: Management for the Year 2000. New York: Morgan Kaufman

Smith, M.(1993), Neural Networks for Statistical Modeling. New York: Van Nostrand Reinhold

Sprauge, R. H., and H. Watson (1994), Decision Support for Management. Upper Saddle River, N.J.: Prentice Hall

Thomsen, Erik (1997), OLAP Solutions: Building Multidimensional Information Systems. New York: Wiley.

Tukey J. W. (1958), "Bias and confidence in Not Quite Large Samples" Annals of Mathematical Statistics ,29:614

Turban, E. (1995), Decision Support Systems and Expert Systems. Upper Saddle River, N.J.: Prentice Hall.

Ulman, J.D. (1989), Principles of Database Systems .Santa Clara, Calif.: Computer Science Press.

Weiss, K. M., and C.A. Kulikuwski (1991), Computer Systems that Learn. New York: Morgan Kaufman.