Imbalanced Multiclass Classification with E.coli Dataset

Πανώριος Μιχαήλ
Πανεπιστήμιο Πειραιά
Τμήμα Ψηφιακών Συστημάτων
Αριθμός Μητρώου: e18127
michaelpanorios@gmail.com

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή κληθήκαμε να εφαρμόσουμε στην πράξη έναν ή περισσότερους αλγόριθμους κατηγοριοποίησης (classification algorithms) σε ένα πραγματικό σύνολο δεδομένων. Η κατηγοριοποίηση του συνόλου δεδομένων θα γίνει με βάση τις μεθόδους κατηγοριοποίησης. Η έμφαση στην παρούσα εργασία είναι στην ορθότητα της μεθοδολογίας που θα ακολουθηθεί για την προετοιμασία των δεδομένων, για την κατασκευή του μοντέλου, και για την αποτίμησή του.

ΕΙΣΑΓΩΓΗ

Σκοπός της εργασίας είναι να δουλέψουμε πάνω σε datasets προκειμένου να εξοικειωθούμε με την κατηγοριοποίηση των δεδομένων. Μέσω αυτής της εργασίας πρόκειται να αποκτήσουμε μια καλύτερη γνώση στην έννοια της κατηγοριοποίησης, στα μετρικά και την αξιολόγηση των δεδομένων.

ΠΕΡΙΓΡΑΦΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Σε αυτό το έργο, θα χρησιμοποιήσουμε ένα τυπικό μη ισορροπημένο σύνολο δεδομένων που αναφέρεται ως σύνολο δεδομένων "Ecoli", που αναφέρεται επίσης ως σύνολο δεδομένων "εντοπισμού πρωτεϊνών". Το σύνολο δεδομένων περιγράφει το πρόβλημα της ταξινόμησης των πρωτεϊνών E.coli χρησιμοποιώντας τις αλληλουχίες αμινοξέων τους στις θέσεις εντοπισμού κυττάρων. ο σύνολο δεδομένων αποτελείται από 336 παραδείγματα πρωτεϊνών Ecoli και κάθε παράδειγμα περιγράφεται χρησιμοποιώντας επτά μεταβλητές εισόδου που υπολογίζονται από την αλληλουχία αμινοξέων πρωτεϊνών. Δηλαδή, προβλέποντας πώς μια πρωτεΐνη θα συνδεθεί με ένα κύτταρο με βάση τη χημική σύνθεση της πρωτεΐνης πριν αναδιπλωθεί. Αγνοώντας το όνομα της ακολουθίας, τα χαρακτηριστικά εισαγωγής περιγράφονται ως εξής: **mcg**: Η μέθοδος της McGeoch για αναγνώριση ακολουθίας σήματος. **gvh**: μέθοδος von Heijne για αναγνώριση ακολουθίας σήματος. **lip**: Βαθμολογία συναίνεσης αλληλουχίας. **chg**: Παρουσία φορτίου λιποπρωτεϊνών. **aac**: βαθμολογία της διακριτικής ανάλυσης της περιεκτικότητας σε αμινοξέα της εξωτερικής μεμβράνης και των περιπλασματικών πρωτεϊνών. **alm1**: βαθμολογία του προγράμματος πρόβλεψης περιοχής μεμβράνης ALOM. ,**alm2**: βαθμολογία του προγράμματος ALOM μετά την εξαίρεση πιθανών περιοχών σήματος που μπορούν να διαχωριστούν από την ακολουθία. Υπάρχουν οκτώ τάξεις που περιγράφονται ως εξής: **cp**: κυτταρόπλασμα, **im**: εσωτερική μεμβράνη χωρίς αλληλουχία, **ρρ**: περίπλασμα **imU**: εσωτερική μεμβράνη, μη διασπώμενη αλληλουχία σήματος, **om**: εξωτερική μεμβράνη **omL**: λιποπρωτεΐνη εξωτερικής μεμβράνης **imL**: λιποπρωτεΐνη εσωτερικής μεμβράνης **imS**: εσωτερική μεμβράνη, διασπώμενη αλληλουχία σήματος Η κατανομή των παραδειγμάτων σε όλες τις τάξεις δεν είναι ίση και σε ορισμένες περιπτώσεις ιδιαίτερα ανισόρροπη.

ΒΗΜΑΤΑ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑΣ

Η έκδοση του συνόλου δεδομένων (έχει αφαιρεθεί η πρώτη στήλη όνομα ακολουθίας) καθώς δεν περιέχονται γενικευμένες πληροφορίες για μοντελοποίηση. Η πληροφορία στο αρχείο ecoli,csv παρουσιάζονται με αυτόν τον τρόπο (παράδειγμα που καλύπτει τις πέντε πρώτες σειρές και παράδειγμα ενός heatmap):

```
0.49,0.29,0.48,0.50,0.56,0.24,0.35,cp
0.07,0.40,0.48,0.50,0.54,0.35,0.44,cp
0.56,0.40,0.48,0.50,0.49,0.37,0.46,cp
0.59,0.49,0.48,0.50,0.52,0.45,0.36,cp
0.23,0.32,0.48,0.50,0.55,0.25,0.35,cp
```

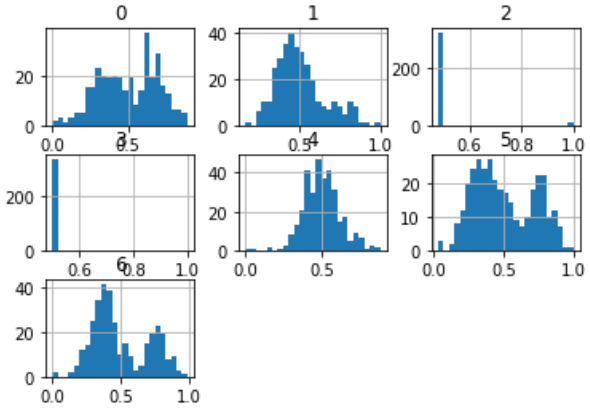
Μπορούμε να δούμε ότι όλες οι μεταβλητές εισόδου εμφανίζονται ως αριθμητικές και οι τιμές που προσδιορίζουν την κλάση είναι τιμές string που θα πρέπει να γίνουν encoding πριν από τη μοντελοποίηση. Το σύνολο δεδομένων μπορεί να φορτωθεί ως

dataframe χρησιμοποιώντας τη συνάρτηση read_csv (), καθορίζοντας τη θέση του αρχείου και το γεγονός ότι δεν υπάρχει heading. Η εκτέλεση του παραδείγματος φορτώνει πρώτα το σύνολο δεδομένων και επιβεβαιώνει τον αριθμό γραμμών και στηλών, οι οποίες είναι 336 σειρές και 7 μεταβλητές εισόδου και 1 μεταβλητή στόχου. Εξετάζοντας την περίληψη κάθε μεταβλητής, φαίνεται ότι οι μεταβλητές έχουν κεντραριστεί, δηλαδή μετατοπίστηκαν ώστε να έχουν μέσο όρο 0,5. Φαίνεται επίσης ότι οι μεταβλητές έχουν κανονικοποιηθεί, που σημαίνει ότι όλες οι τιμές κυμαίνονται μεταξύ περίπου 0 και 1. Τουλάχιστον καμία μεταβλητή δεν έχει τιμές εκτός αυτού του εύρους. Στη συνέχεια, η κατανομή της τάξης συνοψίζεται, επιβεβαιώνοντας τη σοβαρή κλίση στις παρατηρήσεις για κάθε τάξη. Μπορούμε να δούμε ότι η κλάση "cp" κυριαρχεί με το 42% περίπου των παραδειγμάτων και οι τάξεις μειονοτήτων όπως "imS", "imL" και "omL" έχουν περίπου 1 τοις εκατό ή λιγότερο του συνόλου δεδομένων. Ενδέχεται να μην υπάρχουν επαρκή δεδομένα για τη γενίκευση από αυτές τις τάξεις με τις μειονότητες. Ενδέχεται να μην υπάρχουν επαρκή δεδομένα για τη γενίκευση από αυτές τις τάξεις μειονοτήτων.

```
       0      1      2      3      4      5      6
count  336.000  336.000  336.000  336.000  336.000  336.000  336.000
mean   0.500  0.500  0.495  0.501  0.500  0.500  0.500
std    0.195  0.148  0.088  0.027  0.122  0.216  0.209
min    0.000  0.160  0.480  0.500  0.000  0.030  0.000
25%    0.340  0.400  0.480  0.500  0.420  0.330  0.350
50%    0.500  0.470  0.480  0.500  0.495  0.455  0.430
75%    0.662  0.570  0.480  0.500  0.570  0.710  0.710
max    0.890  1.000  1.000  1.000  0.880  1.000  0.990

Class=cp, Count=143, Percentage=42.560%
Class=im, Count=77, Percentage=22.917%
Class=imS, Count=2, Percentage=0.595%
Class=imL, Count=2, Percentage=0.595%
Class=imU, Count=35, Percentage=10.417%
Class=om, Count=20, Percentage=5.952%
Class=omL, Count=5, Percentage=1.488%
Class=pp, Count=52, Percentage=15.476%
```

Μπορούμε επίσης να ρίξουμε μια ματιά στην κατανομή των μεταβλητών εισόδου δημιουργώντας ένα ιστόγραμμα για κάθε μία. Το πλήρες παράδειγμα ιστογραμμάτων όλων των μεταβλητών εισόδου παρατίθεται παρακάτω.



Η διαδικασία πολλαπλής επικύρωσης k-fold παρέχει μια καλή εκτίμηση της απόδοσης του μοντέλου, τουλάχιστον σε σύγκριση με ένα διαχωρισμό single-train test. Θα χρησιμοποιήσουμε το k = 5, που σημαίνει ότι κάθε πτυχή θα περιέχει περίπου 336/5 ή περίπου 67 παραδείγματα, που σημαίνει ότι κάθε πτυχή θα στοχεύει να περιέχει το ίδιο μείγμα παραδειγμάτων ανά τάξη με ολόκληρο το σύνολο δεδομένων εκπαίδευσης για τον κάθε αλγόριθμο που πρόκειται να χρησιμοποιηθεί. Επαναλαμβανόμενο σημαίνει ότι η διαδικασία αξιολόγησης θα πραγματοποιηθεί πολλές φορές για να αποφευχθούν αποτυχημένα αποτελέσματα και να καταγραφεί καλύτερα η διακύμανση του επιλεγμένου μοντέλου. Θα χρησιμοποιήσουμε τρεις επαναλήψεις. Αυτό σημαίνει ότι ένα μονό μοντέλο θα είναι κατάλληλο και θα αξιολογείται 5 * 3 ή 15 φορές και θα αναφέρεται η μέση και τυπική απόκλιση αυτών των διαδρομών. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας την Scikit-Learn RepeatStratifiedKFold. Όλα τα μαθήματα είναι εξίσου σημαντικά. Ως τέτοια, σε αυτήν την περίπτωση, θα χρησιμοποιήσουμε την ακρίβεια ταξινόμησης για την αξιολόγηση μοντέλων. Πρώτον, μπορούμε να ορίσουμε μια συνάρτηση για τη φόρτωση του συνόλου δεδομένων και να χωρίσουμε τις μεταβλητές εισόδου και μεταβλητές εξόδου και να χρησιμοποιήσουμε έναν κωδικοποιητή για να διασφαλίσουμε ότι οι ετικέτες κλάσης αριθμούνται διαδοχικά. Μπορούμε να ορίσουμε μια συνάρτηση για την αξιολόγηση ενός υποψηφίου μοντέλου χρησιμοποιώντας επαναλαμβανόμενη πενταπλή επικύρωση (cross-validation) και, στη συνέχεια, επιστρέφουμε μια λίστα βαθμολογιών που υπολογίζονται στο μοντέλο για κάθε node και επανάληψη. Φορτώντας τα δεδομένα ecoli κρίνεται απαραίτητο να βρούμε την πλειονότητα των κλάσεων και να ακολουθήσουμε μια στρατηγική. Το DummyClassifier μας βοηθά με δύο παραμέτρους που παρέχει το strategy και most_frequent ψάχνοντας την ακρίβεια με accuracy_score για κάποια κλάση από το set ecoli π.χ cp. με την προεπιλεγμένη στρατηγική μας στη συνέχεια αξιολογείται χρησιμοποιώντας k-fold, cross-validation και η μέση και τυπική απόκλιση της ακρίβειας η οποία αναφέρεται ως περίπου 42,6 %. Η βασική γραμμή στην απόδοση καθορίστηκε στο 43,1%. Αυτή η βαθμολογία παρέχει μια βάση για αυτό το σύνολο δεδομένων με το οποίο μπορούν να συγκριθούν όλοι οι άλλοι αλγόριθμοι ταξινόμησης. Η επίτευξη βαθμολογίας πάνω από περίπου 43,1 τοις εκατό υποδηλώνει ότι ένα μοντέλο έχει δεξιότητα σε αυτό το σύνολο δεδομένων και μια βαθμολογία κάτω από αυτήν την τιμή δείχνει ότι το μοντέλο δεν έχει δεξιότητα σε αυτό το σύνολο δεδομένων οπότε δεν συνίσταται να εφαρμοστεί.

ΑΛΓΟΡΙΘΜΟΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Για τους αλγορίθμους κατηγοριοποίησης είναι στοίχημα να δούμε κατά πόσο οι επιλογές μας θα εφαρμόζονται στο υπάρχον σετ δεδομένων. Αξίζει να σημειωθεί ότι θα μπορούσαν να χρησιμοποιηθούν πληθώρα άλλων αλγορίθμων. Μπορεί να είναι βέβαια καλή ιδέα να ελέγξουμε μια σειρά διαφορετικών μη γραμμικών αλγορίθμων σε ένα σύνολο δεδομένων για να δούμε τι είναι αυτό που λειτουργεί καλά και αξίζει περαιτέρω προσοχή και τι όχι. Θα αξιολογήσουμε τα ακόλουθα στο σύνολο δεδομένων E.coli:

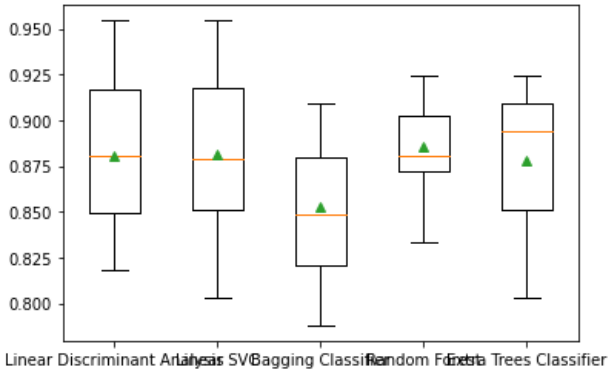
- ***Linear Discriminant Analysis (LDA)***
- ***Support Vector Machine (SVM)***
- ***Bagged Decision Trees (BAG)***

- ***Random Forest (RF)***
- ***Extra Trees (ET)***

ΜΕΘΟΔΟΛΟΓΙΑ

Θα χρησιμοποιήσουμε ως επί το πλείστον προεπιλεγμένες παραμέτρους μοντέλου, με εξαίρεση τον αριθμό δέντρων στους αλγόριθμους συνόλου, τους οποίους θα ορίσουμε σε μια προεπιλογή 1.000. Θα ορίσουμε κάθε μοντέλο με τη σειρά του και θα τα προσθέσουμε σε μια λίστα, ώστε να τα αξιολογούμε διαδοχικά. Σε αυτήν την περίπτωση, μπορούμε να δούμε ότι όλοι οι αλγόριθμοι που δοκιμάστηκαν έχουν δεξιότητα, επιτυγχάνοντας ακρίβεια πάνω από το ενδέδειγμένο 43,1%. Τα αποτελέσματα υποδηλώνουν ότι οι περισσότεροι αλγόριθμοι τα πάνε καλά σε αυτό το σύνολο δεδομένων και ότι ίσως τα σύνολα των δέντρων απόφασης αποδίδουν καλύτερα με τα Extra Trees να επιτυγχάνουν 87,8% ακρίβεια και το Random Forest να επιτυγχάνει 88,6% ακρίβεια. Δημιουργείται ένας αριθμός που δείχνει ένα πλαίσιο και ένα γράφημα για το δείγμα αποτελεσμάτων κάθε αλγορίθμου. Το πλαίσιο δείχνει το μέσο 50 % των δεδομένων, η πορτοκαλί γραμμή στο μέσο κάθε πλαισίου δείχνει τη μέση τιμή του δείγματος και το πράσινο τρίγωνο σε κάθε πλαίσιο δείχνει τη μέση τιμή του δείγματος. Μπορούμε να δούμε ότι οι κατανομές των βαθμολογιών για τα σύνολα των δέντρων αποφάσεων συγκεντρώθηκαν χωριστά από τους άλλους αλγορίθμους που δοκιμάστηκαν. Στις περισσότερες περιπτώσεις, η διαμέσος και ο μέσος όρος είναι κοντά στο μέσο όρο, υποδηλώνοντας μια κάπως συμμετρική κατανομή των βαθμολογιών που μπορεί να υποδηλώνουν ότι τα μοντέλα είναι σταθερά. Το αποτέλεσμα που παίρνουμε από το κώδικα είναι:

```
>Linear Discriminant Analysis 0.881 (0.041)
>Linear SVC 0.882 (0.040)
>Bagging Classifier 0.852 (0.036)
>Random Forest 0.886 (0.025)
>Extra Trees Classifier 0.878 (0.037)
```

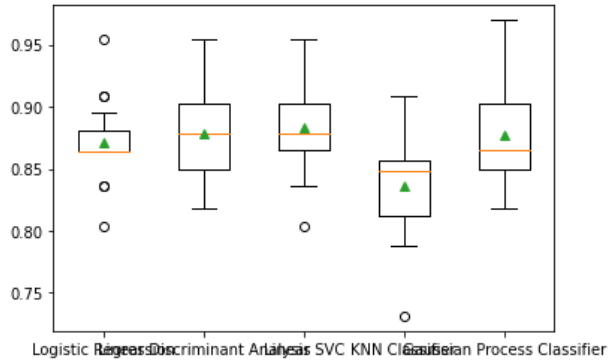


Μπορούμε να δοκιμάσουμε τον αλγόριθμο SMOTE που εφαρμόζεται σε όλους εκτός από την κατηγορία πλειοψηφίας (cp) που οδηγεί σε αύξηση της απόδοσης. Γενικά, το SMOTE δεν φαίνεται να βοηθά τα σύνολα των δέντρων αποφάσεων, επομένως θα αλλάξουμε το σύνολο των αλγορίθμων που δοκιμάστηκαν στα ακόλουθα

- ***Multinomial Logistic Regression (LR)***
- ***Linear Discriminant Analysis (LDA)***
- ***Support Vector Machine (SVM)***
- ***k-Nearest Neighbors (KNN)***
- ***Gaussian Process (GP)***

Μπορούμε να χρησιμοποιήσουμε την εφαρμογή SMOTE από τη βιβλιοθήκη imbalanced-learn και έναν pipeline από την ίδια βιβλιοθήκη για να εφαρμόσουμε πρώτα το SMOTE στο σύνολο δεδομένων εκπαίδευσης και στη συνέχεια να ταιριάξουμε ένα συγκεκριμένο μοντέλο ως μέρος της διαδικασίας cross-validate. Το SMOTE θα συνθέσει νέα παραδείγματα χρησιμοποιώντας k-πλησιέστερους γείτονες στο σύνολο δεδομένων εκπαίδευσης, όπου από προεπιλογή, το k είναι → 5. Αυτό είναι πολύ μεγάλο για ορισμένες από τις τάξεις στο σύνολο δεδομένων μας. Επομένως, θα δοκιμάσουμε μια τιμή k είναι → 2. Σε αυτήν την περίπτωση, μπορούμε να δούμε ότι το LDA με το SMOTE είχε ως αποτέλεσμα μια μικρή πτώση από 88,6% σε περίπου 87,9%, ενώ η SVM με την SMOTE σημείωσε μικρή αύξηση από περίπου 88,2% σε περίπου 88,4%. Το SVM φαίνεται επίσης να είναι η μέθοδος με την καλύτερη απόδοση όταν χρησιμοποιείτε το SMOTE σε αυτήν την περίπτωση, αν και δεν επιτυγχάνεται βελτίωση σε σύγκριση με το τυχαίο δάσος στην προηγούμενη ενότητα.

```
>Logistic Regression 0.871 (0.035)
>Linear Discriminant Analysis 0.879 (0.041)
>Linear SVC 0.884 (0.039)
>KNN Classifier 0.836 (0.044)
>Gaussian Process Classifier 0.877 (0.038)
```



Μπορούμε να δούμε ότι η LR έχει μια σειρά επιδόσεις με υψηλές τιμές 87 τοις εκατό, κάτι που είναι αρκετά ενδιαφέρον. Μπορεί να υποδηλώνει ότι το LR θα μπορούσε να αποδώσει καλύτερα. Μπορούμε να χωρέσουμε σε ένα τελικό μοντέλο και να το χρησιμοποιήσουμε για να κάνουμε προβλέψεις σε μεμονωμένες σειρές δεδομένων. Θα χρησιμοποιήσουμε το μοντέλο Random Forest ως το τελικό μας μοντέλο που πέτυχε ακρίβεια ταξινόμησης περίπου 88,6%. Για να το αποδείξουμε αυτό, μπορούμε να χρησιμοποιήσουμε μοντέλο προσαρμογής για να κάνουμε κάποιες προβλέψεις ετικετών για μερικές περιπτώσεις όπου γνωρίζουμε το αποτέλεσμα. Η εκτέλεση του παραδείγματος ταιριάζει πρώτα με το μοντέλο σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης. Στη συνέχεια, το μοντέλο προσαρμογής χρησιμοποιείται για την πρόβλεψη της ετικέτας για ένα παράδειγμα που λαμβάνεται από καθεμία από τις έξι κατηγορίες. Μπορούμε να δούμε ότι η σωστή

ετικέτα κατηγορίας προβλέπεται για καθένα από τα επιλεγμένα παραδείγματα. Ωστόσο, κατά μέσο όρο, αναμένουμε ότι 1 στις 10 προβλέψεις θα είναι λάθος και αυτά τα σφάλματα ενδέχεται να μην κατανέμονται εξίσου σε όλες τις τάξεις. Ωστόσο στα δικά μας δεδομένα τα αποτελέσματα είναι σωστά με 100% επιτυχία.

```
>Predicted=cp (expected cp)
>Predicted=im (expected im)
>Predicted=imU (expected imU)
>Predicted=om (expected om)
>Predicted=omL (expected omL)
>Predicted=pp (expected pp)
```

ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΤΡΙΚΩΝ

Επιλογή πρώτων classifiers μη-γραμμικών για να δούμε τι ταιριάζει καλύτερα στο σύνολο μας. (Classifier: Accuracy & SD)

```
>Linear Discriminant Analysis 0.881 (0.041)
>Linear SVC 0.882 (0.040)
>Bagging Classifier 0.852 (0.036)
>Random Forest 0.886 (0.025)
>Extra Trees Classifier 0.878 (0.037)
```

Επιλογή δεύτερων classifiers που εφαρμόζονται στο SMOTE για να δούμε τι ταιριάζει καλύτερα στο σύνολο μας. . (Classifier: Accuracy & SD)

```
>Logistic Regression 0.871 (0.035)
>Linear Discriminant Analysis 0.879 (0.041)
>Linear SVC 0.884 (0.039)
>KNN Classifier 0.836 (0.044)
>Gaussian Process Classifier 0.877 (0.038)
```

ΒΙΒΛΙΟΓΡΑΦΗΚΕΣ ΠΗΓΕΣ

[1] https://analyticsindiamag.com/7-types-classification-algorithms/?fbclid=IwAR3MXUOwk9JAsB8KU0ntSMBzKhOoI70qBsqBrHLKjActdrs77VJDii7ayMU

[2] https://www.kaggle.com/kannanaikkal/balancing-ecoli-data-set-random-over-sampling?fbclid=IwAR36-2S-qmjYgfBmOpEmiYzw-gaEKE316VJq7bibq44wehwDUxyjFd39vXs

[3] https://www.kaggle.com/kannanaikkal/classification-on-ecoli-data-set-balanced?fbclid=IwAR1ax0wV2M95c-Mt6lh4mZGMqRb4zg--HtpcSyazzktnUwzoXnuK_o0_OHo.

[4] Εισαγωγή στην Εξόρυξη Δεδομένων.