Τεχνολογίες Ευφυών Συστημάτων και Ρομποτική

2η Εργασία (CLIPS)
Credit Approval

Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Μονοπάτης Δημήτριος
ΑΜ: 1040546 (Παλαιός ΑΜ: 4776) - Επί διπλώματι
monopatis@ceid.upatras.gr

7 Ιουλίου 2017

Χρήση του Weka 3.8.1 σε GNU/Linux Debian και του CLIPS Rule Based Programming Language 6.30 IDE 64 bit.
Η αναφορά είναι γραμμένη σε LaTeX .
Η κωδικοποίηση των αρχείων είναι UTF-8, με χρήση για αλλαγή νέων γραμμών του line feed (LF).

I. Περιγραφή του προβλήματος

I. Πρόβλημα, παράμετροι, κλάσεις εξόδου

Με βάση τον αριθμό μητρώου επιλέχθηκε το νούμερο 6. (Credit Approval) δηλαδή "Έγκριση Πίστωσης". Το πρόβλημα είναι η έγκριση (+) ή όχι (-) αιτήσεων για παροχή πιστωτικής κάρτας σε πελάτες χρηματοπιστωτικού ιδρύματος. Τα ονόματα των παραμέτρων και τα ονόματα των τιμών τους, έχουν αλλάξει σε κάτι ακατανόητο για την προστασία του απορρήτου.
Οι παράμετροι εισόδου είναι οι A1 μέχρι A15 με τιμές:
A1: b, a.
A2: συνεχής.
A3: συνεχής.
A4: u, y, l, t.
A5: g, p, gg.
A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7: v, h, bb, j, n, z, dd, ff, o.

A8: συνεχής.
A9: t, f.
A10: t, f.
A11: συνεχής.
A12: t, f.
A13: g, p, s.
A14: συνεχής.
A15: συνεχής. Και η κλάση εξόδου είναι η A16 με τιμές:
A16: +,-
Το πλήθος των παραδειγμάτων είναι 690.

II. Προεπεξεργασία συνόλου δεδομένων

Με βάση το αρχείο lisp που παρατίθεται έχουμε τα ονόματα κάποιων παραμέτρων και τις δυνατές τιμές τους, όχι όμως με την σειρά που μας δίνονται, έτσι δεν μπορούμε με σιγουριά να συμπεράνουμε τις αντιστοιχίες. Υποθέτω ότι το A9 είναι το γένος (t:female f:male).
Οι άλλοι παράμετροι είναι:
jobless
purchase_item με τιμές stereo, pc κλπ
unmarried
problematic_region
age
deposit
monthly_payment
numb_of_months
numb_of_years_in_company

Στο αρχικό σύνολο δεδομένων (ΣΔ) έγινε αφαίρεση της παραμέτρους A1 διότι υποθέτω ότι αναφέρεται σε δύο διαφορετικά σετ παραδειγμάτων που συγχωνεύτηκαν και έτσι δεν παίζει ρόλο. Κατόπιν έγινε αφαίρεση των παραδειγμάτων που έχουν ελλιπείς τιμές και το πλήθος τους έγινε 646. Και τέλος αφαιρέθηκε η παράμετρος A6 διότι είχε πολλές διακριτές τιμές που μεγάλωναν την πολυπλοκότητα άσκοπα. Για τον διαχωρισμό του (ΣΔ) εφαρμόστηκε πρώτα ένα τυχαίο ανακάτεμα με χρήση φίλτρου του Weka (Randomize) και στην συνέχεια με χρήση του φίλτρου Remove σε ποσοστά 2/3 για το training και 1/3 για το test.
Ακόμη δημιουργηθήκαν τα αρχεία εισόδου για το πρόγραμμα CLIPS με αντικατάσταση του χαρακτήρα ” , ” με κενό.

II. Δημιουργία Δέντρων Αποφάσεων

I. Περιγραφή εξαγωγής δέντρων αποφάσεων μέσω WEKA

Στο Weka με τον αλγόριθμο J48 ανάλογα και με τις παραμέτρους του, παράγεται το δέντρο απόφασης για το training set (ΣΕΚ - σύνολο εκπαίδευσης).

II. Περιγραφή προσπαθειών βελτίωσης δέντρων

Μετά από δοκιμές, οι παράμετροι του J48 για το δεύτερο καλύτερο δέντρο αποφάσεων ήταν:

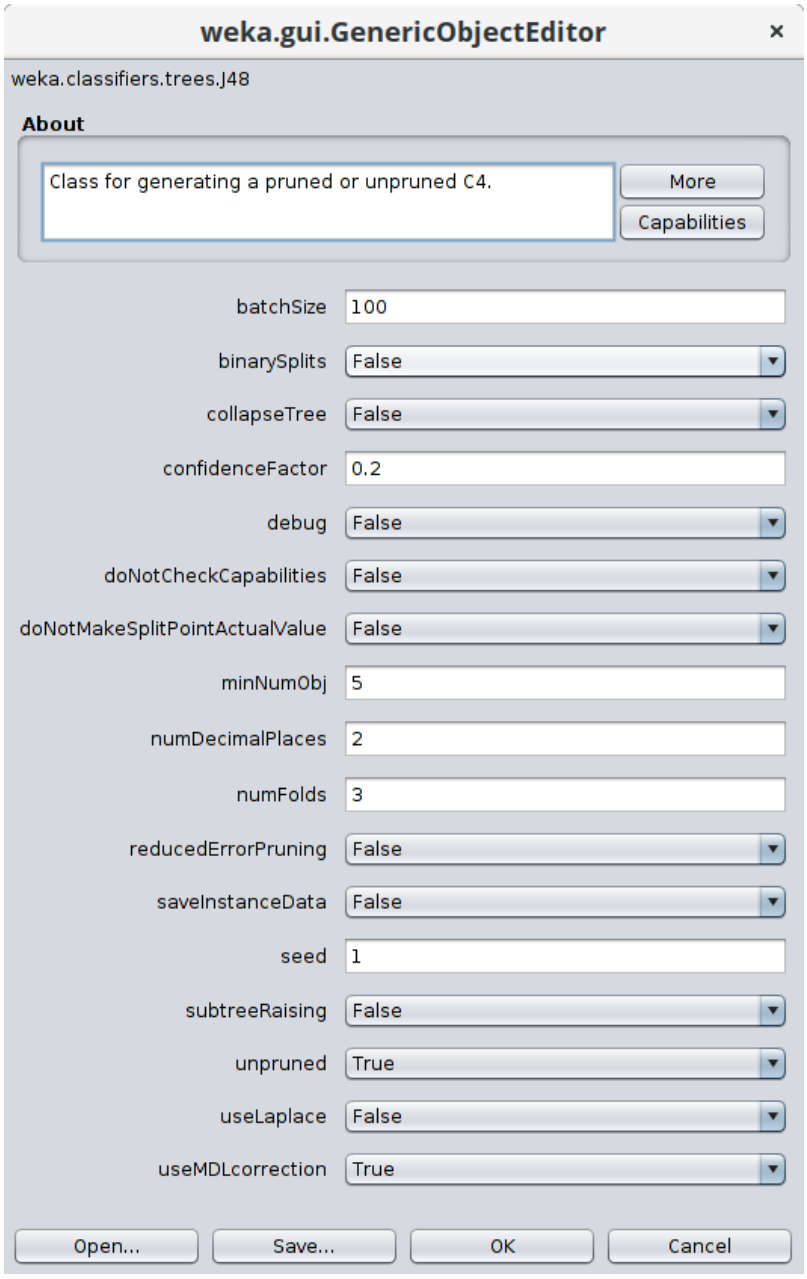


Figure 1: *Σχήμα 1*

Το δέντρο που κατασκευάστηκε είναι

```
A9 = t
| A10 = t
| | A11 <= 3
| | | A15 <= 501
| | | | A4 = u
| | | | | A7 = v
| | | | | | A11 <= 2
| | | | | | | A12 = t: + (5.0/2.0)
| | | | | | | A12 = f: + (6.0/2.0)
| | | | | | A11 > 2: + (9.0/1.0)
| | | | | A7 = h
| | | | | | A2 <= 32.83: + (7.0)
| | | | | | A2 > 32.83: + (5.0/2.0)
| | | | | A7 = bb: + (1.0)
| | | | | A7 = j: + (0.0)
| | | | | A7 = n: + (0.0)
| | | | | A7 = z: + (2.0)
| | | | | A7 = dd: + (0.0)
| | | | | A7 = ff: + (2.0/1.0)
| | | | | A7 = o: + (0.0)
| | | | A4 = y: + (7.0/3.0)
| | | | A4 = l: + (0.0)
| | | | A4 = t: + (0.0)
| | | A15 > 501: + (20.0)
| | A11 > 3
| | | A7 = v
| | | | A2 <= 22.5: + (6.0/2.0)
| | | | A2 > 22.5: + (40.0)
| | | A7 = h: + (22.0)
| | | A7 = bb
| | | | A12 = t: + (8.0/1.0)
| | | | A12 = f: + (7.0)
| | | A7 = j: + (1.0)
| | | A7 = n: + (0.0)
| | | A7 = z: + (2.0)
| | | A7 = dd: + (0.0)
| | | A7 = ff: + (1.0)
| | | A7 = o: + (0.0)
| A10 = f
| | A14 <= 70
| | | A7 = v
| | | | A8 <= 2: + (5.0/1.0)
| | | | A8 > 2: + (6.0)
| | | A7 = h: + (6.0)
| | | A7 = bb: + (3.0/1.0)
| | | A7 = j: + (0.0)
| | | A7 = n: + (0.0)
| | | A7 = z: + (1.0)
```

```
| | | A7 = dd: + (1.0)
| | | A7 = ff: + (0.0)
| | | A7 = o: + (0.0)
| | A14 > 70
| | | A15 <= 109
| | | | A4 = u
| | | | | A3 <= 2.665
| | | | | | A12 = t: - (8.0)
| | | | | | A12 = f
| | | | | | | A14 <= 290: - (5.0)
| | | | | | | A14 > 290: + (6.0/3.0)
| | | | | A3 > 2.665
| | | | | | A12 = t: - (9.0/4.0)
| | | | | | A12 = f: + (8.0/1.0)
| | | | A4 = y
| | | | | A7 = v: - (6.0)
| | | | | A7 = h: - (5.0/2.0)
| | | | | A7 = bb: - (0.0)
| | | | | A7 = j: - (0.0)
| | | | | A7 = n: - (0.0)
| | | | | A7 = z: - (0.0)
| | | | | A7 = dd: - (0.0)
| | | | | A7 = ff: - (1.0)
| | | | | A7 = o: - (0.0)
| | | | A4 = l: - (0.0)
| | | | A4 = t: - (0.0)
| | | A15 > 109
| | | | A12 = t: + (7.0/2.0)
| | | | A12 = f: + (5.0)
A9 = f
| A3 <= 0.165
| | A12 = t: + (6.0/3.0)
| | A12 = f
| | | A4 = u: + (6.0/3.0)
| | | A4 = y: - (6.0)
| | | A4 = l: - (0.0)
| | | A4 = t: - (0.0)
| A3 > 0.165
| | A13 = g
| | | A4 = u
| | | | A10 = t: - (34.0)
| | | | A10 = f
| | | | | A7 = v
| | | | | | A12 = t: - (21.0/1.0)
| | | | | | A12 = f: - (37.0/3.0)
| | | | | A7 = h: - (9.0)
| | | | | A7 = bb
| | | | | | A2 <= 37.75: - (5.0)
| | | | | | A2 > 37.75: - (5.0/1.0)
| | | | | A7 = j: - (1.0)
```

| | | | | A7 = n: - (0.0)
| | | | | A7 = z: - (1.0)
| | | | | A7 = dd: - (1.0)
| | | | | A7 = ff: - (6.0)
| | | | | A7 = o: - (0.0)
| | | A4 = y: - (52.0)
| | | A4 = l: - (0.0)
| | | A4 = t: - (0.0)
| | A13 = p: + (1.0)
| | A13 = s: - (18.0) Number of Leaves : 77
Size of the tree : 109

Μπορείτε να δείτε την γραφική αναπαράσταση και στο φάκελο Trees

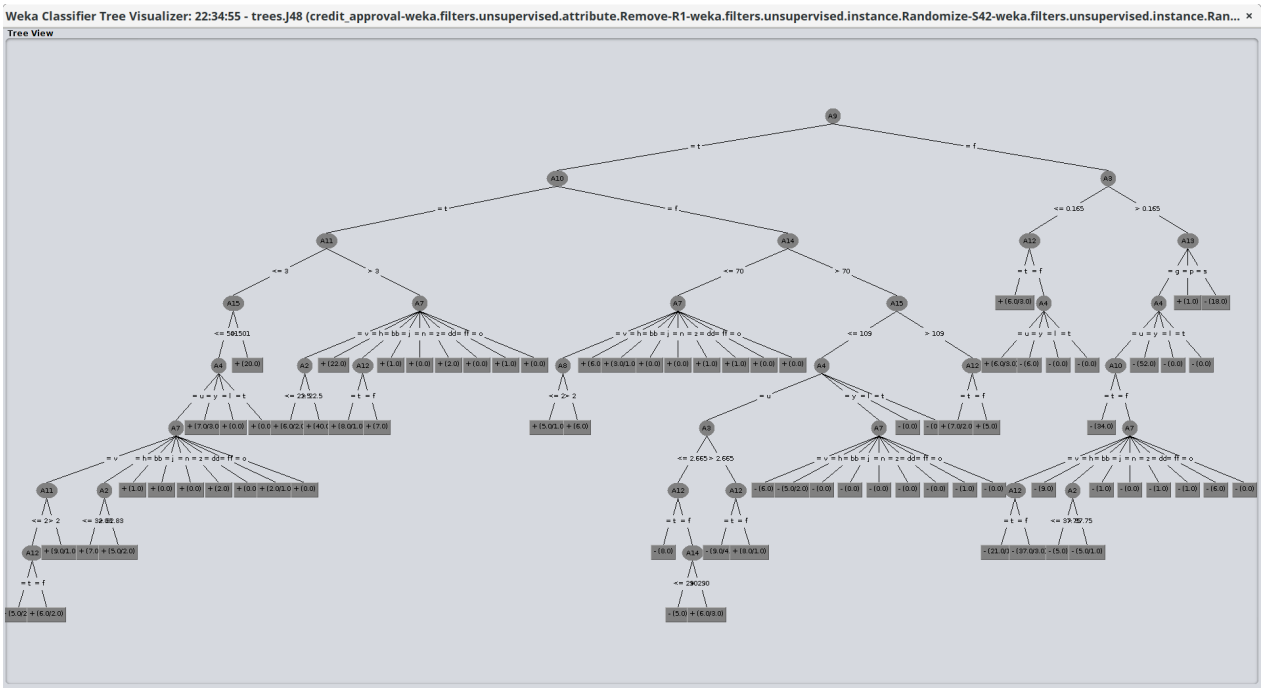
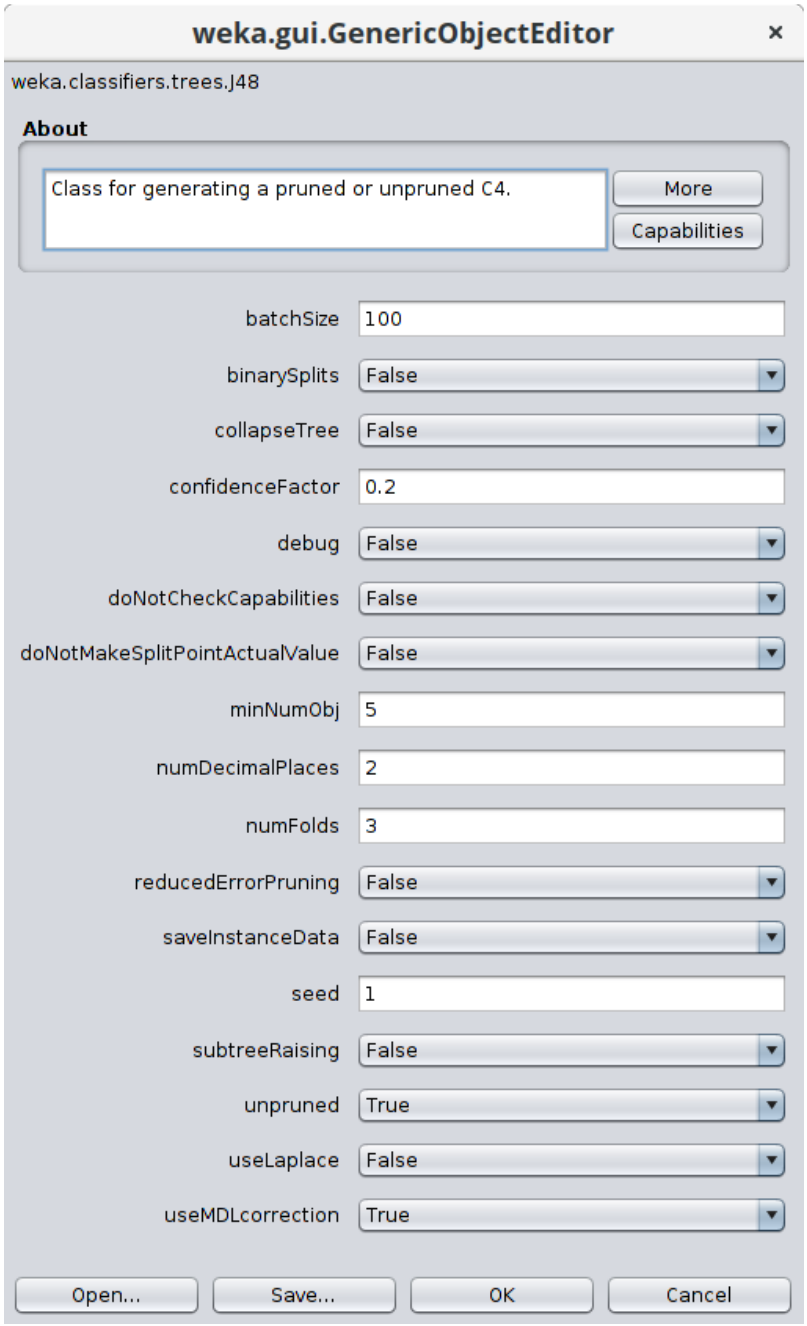


Figure 2: *tree_2*

Οι παράμετροι του J48 για το καλύτερο δέντρο αποφάσεων ήταν:

-L -U -M 5



Και το καλύτερο δέντρο αποφάσεων:

A9 = t
| A10 = t: + (151.0/14.0)
| A10 = f
| | A14 <= 70: + (22.0/2.0)
| | A14 > 70

```
| | | A15 <= 109
| | | | A4 = u
| | | | | A3 <= 2.665: - (19.0/3.0)
| | | | | A3 > 2.665
| | | | | | A12 = t: - (9.0/4.0)
| | | | | | A12 = f: + (8.0/1.0)
| | | | A4 = y: - (12.0/2.0)
| | | | A4 = l: - (0.0)
| | | | A4 = t: - (0.0)
| | | A15 > 109: + (12.0/2.0)
A9 = f
| A3 <= 0.165: - (18.0/6.0)
| A3 > 0.165
| | A13 = g: - (172.0/5.0)
| | A13 = p: + (1.0)
| | A13 = s: - (18.0)
Number of Leaves : 13
Size of the tree : 22
```

Μπορείτε να δείτε την γραφική αναπαράσταση και στο φάκελο Trees

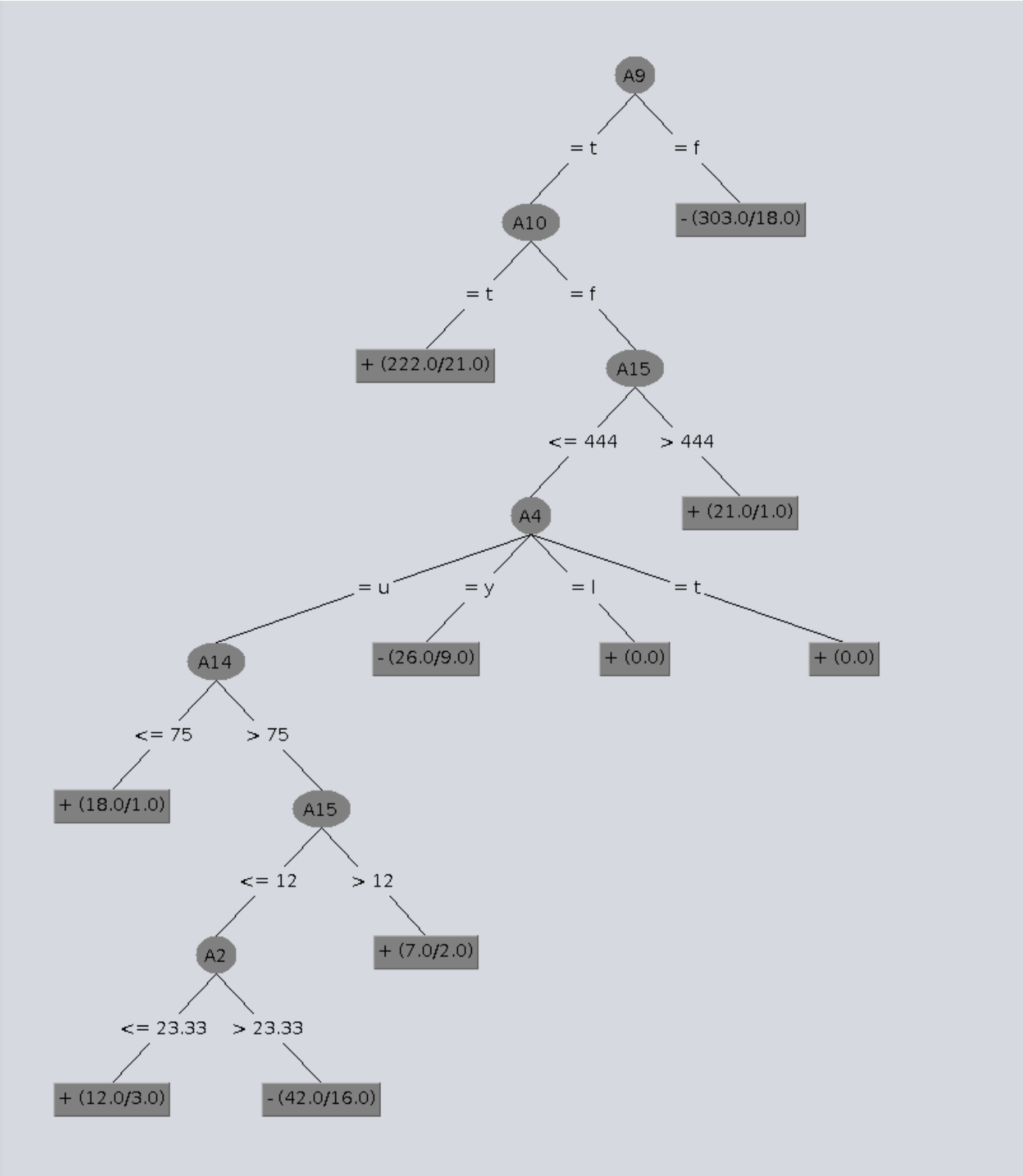


Figure 3: *tree_1*

III. Αποτελέσματα μετρικών στο σύνολο εκπαίδευσης

Για το δεύτερο καλύτερο έχουμε:
Ορθότητα: 91,1765%
Ακρίβεια: 0,868
Ευαισθησία: 0,943589743589743
Εξειδίκευση: 0,88663967611336

Με
TP = 184
FN = 11
FP = 28
TN = 219

Για το πρώτο καλύτερο έχουμε:
Ορθότητα: 91,1765%
Ακρίβεια: 0,902
Ευαισθησία: 0,897435897435897
Εξειδίκευση: 0,923076923076923

Με
TP = 175
FN = 20
FP = 19
TN = 228

IV. Αποτελέσματα μετρικών στο σύνολο ελέγχου

Για το δεύτερο καλύτερο έχουμε:
Ορθότητα: 85,0679%
Ακρίβεια: 0,845
Ευαισθησία: 0,836538461538462
Εξειδίκευση: 0,863247863247863

Με
TP = 87
FN = 17
FP = 16
TN = 101

Για το πρώτο καλύτερο έχουμε:
Ορθότητα: 86,4253%

Ακρίβεια: 0,878
Ευαισθησία: 0,826923076923077
Εξειδίκευση: 0,897435897435897

Με
TP = 86
FN = 18
FP = 12
TN = 105

III. Υλοποίηση σε CLIPS

I. Εξαγωγή κανόνων από WEKA

Οι κανόνες σύμφωνα με το καλύτερο δέντρο αποφάσεων εφαρμοσμένο για το σύνολο ελέγχου είναι:

A9 = t
| A10 = t: + (151.0/14.0) **r1**
| A10 = f
| | A14 <= 70: + (22.0/2.0) **r2**
| | A14 > 70
| | | A15 <= 109
| | | | A4 = u
| | | | | A3 <= 2.665: - (19.0/3.0) **r3**
| | | | | A3 > 2.665
| | | | | | A12 = t: - (9.0/4.0) **r4**
| | | | | | A12 = f: + (8.0/1.0) **r5**
| | | | A4 = y: - (12.0/2.0) **r6**
| | | | A4 = l: - (0.0) **r7**
| | | | A4 = t: - (0.0) **r8**
| | | A15 > 109: + (12.0/2.0) **r9**
A9 = f
| A3 <= 0.165: - (18.0/6.0) **r10**
| A3 > 0.165
| | A13 = g: - (172.0/5.0) **r11**
| | A13 = p: + (1.0) **r12**
| | A13 = s: - (18.0) **r13**

Με **σκούρα** γράμματα τα ονόματά τους.

II. Περιγραφή υλοποίησης των κανόνων σε CLIPS

Γίνεται φόρτωση των παραδειγμάτων ως γεγονότα με την assert στο batch αρχείο crx.BAT από το αρχείο που αναφέρεται στην γραμμή 38. Οι κανόνες φορτώνονται από το αρχείο crx.clp που γίνεται και ο υπολογισμός των μετρικών αξιολόγησης.

III. Αλλαγές και «ρύθμιση» του ΕΣ με βάση το σύνολο εκπαίδευσης

Λόγω πολύ καλών αποτελεσμάτων δεν έγινε κάποια αλλαγή.

IV. Αποτελέσματα μετρικών στο σύνολο εκπαίδευσης

Accuracy 0.911904761904762
Sensitivity 0.901162790697674
Specificity 0.919354838709677
Precision 0.885714285714286

V. Αποτελέσματα μετρικών στο σύνολο ελέγχου

Accuracy 0.867924528301887
Sensitivity 0.887640449438202
Specificity 0.853658536585366
Precision 0.814432989690722

IV. Βιβλιογραφία

ΕΥΦΥΗΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ, Ιωάννης Χατζηλυγερούδης, 2013
ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ ΔΗΜΙΟΥΡΓΙΑΣ ΕΜΠΕΙΡΩΝ ΣΥΣΤΗΜΑΤΩΝ ΜΕ ΚΑΝΟΝΕΣ ΓΙΑ ΠΡΟΒΛΗΜΑΤΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΑΠΟ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ, Τζετζούμης Ευάγγελος, 2012