Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων και
Μηχανική Μάθηση

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ»

ΤΕΧΝΙΚΕΣ ΜΕΤΑΜΑΘΗΣΗΣ ΣΕ ΛΙΓΑ ΔΕΔΟΜΕΝΑ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΡΗΣΤΟΣ, Α. ΣΠΥΡΟΠΟΥΛΟΣ

Επιβλέπων: Γεώργιος Στάμου
Αναπληρωτής Καθηγητής, Ε.Μ.Π.
Συνεπίβλεψη: Παρασκευή Τζούβελη
Προσωπικό Ε.Ε.Δ.Ι.Π., Ε.Μ.Π.

Αθήνα, Μάρτιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων και
Μηχανική Μάθηση

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ»

ΤΕΧΝΙΚΕΣ ΜΕΤΑΜΑΘΗΣΗΣ ΣΕ ΛΙΓΑ ΔΕΔΟΜΕΝΑ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΡΗΣΤΟΣ, Α. ΣΠΥΡΟΠΟΥΛΟΣ

Επιβλέπων: Γεώργιος Στάμου
Αναπληρωτής Καθηγητής, Ε.Μ.Π.
Συνεπίβλεψη: Παρασκευή Τζούβελη
Προσωπικό Ε.Ε.Δ.Ι.Π., Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Μαρτίου 2021
Αθήνα, Μάρτιος 2021

Χρήστος, Α. Σπυρόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Η ραγδαία αύξηση του όγκου των δεδομένων τα τελευταία χρόνια, οδήγησε στην άμεση αύξηση της χρήσης βαθιών νευρωνικών δικτύων για την εξαγωγή συμπερασμάτων σε διάφορα προβλήματα. Τα βαθιά νευρωνικά δίκτυα, σε περιπτώσεις προβλημάτων με μεγάλο όγκο δεδομένων είναι πολύ καλοί ταξινομητές και δίνουν λύσεις σε προβλήματα παλινδρόμησης ή βελτιστοποίησης. Αντίθετα, δοκιμαζόμενα σε προβλήματα με λιγότερα δεδομένα απέτυχαν να γενικεύσουν. Η αδυναμία τους αυτή οφείλεται στην έλλειψη δεδομένων, αφού η επιτυχία τους βασίζεται στην ύπαρξη μεγάλων όγκων δεδομένων που με την πολυπλοκότητα τους μπορούν να εξάγουν και να συγκεντρώσουν όλη την πληροφορία που διαθέτουν τα δεδομένα.

Η δημιουργία τεχνικών εξαγωγής συμπερασμάτων με λίγα δεδομένα ήταν απαραίτητη για την αντιμετώπιση τέτοιου είδους προβλημάτων. Η μάθηση με λίγα δεδομένα (Few shot Learning) καταφέρνει να καλύψει αυτό το κενό επιτυγχάνοντας εξαιρετικές αποδόσεις σε προβλήματα με λίγα δεδομένα.

Στην παρούσα διπλωματική εργασία, θα γίνει σύνδεση της Μπεϋζιανής Στατιστικής με τεχνικές μάθησης με λίγα δεδομένα. Ο στόχος της διπλωματικής είναι να χρησιμοποιηθούν τυχαίες μεταβλητές υπό Μπεϋζιανό πλαίσιο, αξιοποιώντας τις δυνατότητες που έχει η μάθηση με λίγα δεδομένα, οι οποίες να συνεισφέρουν θετικά σε προβλήματα ταξινόμησης εικόνων. Τα πειράματα της εργασίας αφορούν την ταξινόμηση εικόνων από εννέα διαφορετικά σετ δεδομένων και την προσαρμογή του προβλήματος σε ρεαλιστικές συνθήκες. Οι ρεαλιστικές συνθήκες αναφέρονται στην ύπαρξη ανισορροπίας μεταξύ των δεδομένων των κλάσσεων και των δεδομένων του κάθε προβλήματος που θα εκπαιδευτεί.

Κάτα την εφαρμογή του πειράματος θα γίνει σύγκριση τριών διαφορετικών τεχνικών μάθησης με λίγα δεδομένα. Οι τεχνικές αυτές θα έχουν παρόμοια αρχιτεκτονική και η μία εκ των οποίων θα χρησιμοποιεί κάτα την κατασκευή της τυχαίες μεταβλητές, οι οποίες θα επηρεάζουν τις αρχικές συνθήκες των παραμέτρων του προβλήματος. Οι συνθήκες αυτές θα διαφέρουν για κάθε κλάση του προβλήματος ανάλογα με το μέγεθος της και θα είναι διαφορετικές για το κάθε πρόβλημα ανάλογα με το μέγεθος του.

Λέξεις Κλειδιά

Μάθηση με λίγα δεδομένα, Μπεϋζιανή Στατιστική, Μετα-Μάθηση, Μετρική-Μάθηση, Νευρωνικά Δίκτυα, Τυχαίες Μεταβλητές, Μετα-Μάθηση σε Διαφορετικά Πεδία, Μπεϋζιανή Προσαρμοσμένη Μεταμάθηση

Abstract

The rapid increase in the data volume in recent years, has led to an immediate increase in the use of deep neural networks. Deep neural networks, in cases of problems with a large amount of data, are very good classifiers and provide solutions to regression or optimization problems. Deep neural networks through their complexity can extract and collect all the information available in the data. In contrast, tested on problems with few data failed to generalize. This weakness is due to the lack of data, since their success is based on the existence of large volumes of data.

The creation of inferential techniques with few data was necessary to address such problems. Few shot Learning manages to fill this gap by achieving excellent performance on few data problems.

In the present dissertation, Bayesian Statistics will be linked to learning techniques with few data. The aim of the dissertation is to use random variables under the Bayesian framework, utilizing the possibilities of learning with few data, which contribute positively to image classification problems. The experiments of the work concern the classification of images from nine different data sets and the adaptation of the problem to realistic conditions. Realistic conditions refer to the existence of an imbalance between the data of the classes and the data of each problem to be trained.

During the implementation of the experiment, three different learning techniques will be compared with few data. These techniques will have a similar architecture and one of them will use random variables during its construction, which will affect the initial conditions of the problem parameters. These conditions will vary for each class of problem depending on its size and will be different for each problem depending on its size.

Key words

Few Shot Learning, Bayesian Statistic, Meta Learning, Metric Learning, Neural Networks, Random Variables, Model Agnostic Meta Learning, Bayesian Task Adaptive Meta Learning

Ευχαριστίες

Πρώτα απ΄ όλα, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή αυτής της διπλωματικής κ. Στάμου Γεώργιο, για την ευκαιρία που μου έδωσε, να ασχοληθώ με το συγκεκριμένο θέμα, αλλά και για την έμπνευση και το ενδιαφέρον που μου καλλιέργησε κατά τη διάρκεια των σπουδών μου.

Ιδιαίτερες ευχαριστίες θα ήθελα να αποδώσω στην Ερευνήτρια κα. Τζούβελη Παρασκευή για την καθοδήγηση της, τη διαρκή και άμεση στήριξη της. Το ενδιαφέρον της για το θέμα της διπλωματικής εργασίας, μου έδωσε σημαντικό κίνητρο για την επιτυχή ολοκλήρωσή της.

Ακόμα, θα ήθελα να ευχαριστήσω όλη την ομάδα των συμμετεχόντων καθηγητών στο Μεταπτυχιακό Πρόγραμμα της Επιστήμης Δεδομένων και Μηχανικής Μάθησης, για τη συνεχή βοήθεια που μου παρείχαν, η οποία συνέβαλε ουσιαστικά στην ομαλή διεξαγωγή της φοίτησης μου στο πρόγραμμα.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που στάθηκαν δίπλα μου και συνέβαλαν στο απαιτητικό αυτό διάστημα των σπουδών μου.

Χρήστος Α. Σπυρόπουλος,

Αθήνα, 1 Μαρτίου 2021

Περιεχόμενα

Κατάλογος σχημάτων

1 Εισαγωγή

Η Τεχνητή Νοημοσύνη είναι ένας τομέας της πληροφορικής που έχει ως στόχο τη δημιουργία δικτύων στα πρότυπα των νευρώνων του ανθρώπινου εγκεφάλου. Τα δίκτυα αυτά αποτελούνται από έναν αριθμό νευρώνων οι οποίοι χρησιμοποιούν τα χαρακτηριστικά του προβλήματος και προσπαθούν να εξάγουν συμπέρασμα για την εξαρτημένη μεταβλητή του προβλήματος. Αυτά τα δίκτυα ονομάζονται νευρωνικά.

Τα νευρωνικά δίκτυα χρησιμοποιούν την τεχνική μετάδοσης γνώσης. Μια τεχνική που χρησιμοποιεί και ο ανθρώπινος εγκέφαλος. Τα νευρωνικά δίκτυα προσπαθούν να προσομοιώσουν τον ανθρώπινο εγκέφαλο γεγονός που μας οδηγεί στο συμπέρασμα ότι η κατασκευή τέτοιων αλγορίθμων είναι μια λογική επιλογή. Επιπλέον, η μετάδοση γνώσης είναι ένα από τα εργαλεία που είναι στενά συνδεδεμένο με την έννοια της Γενικής Τεχνητής νοημοσύνης (Artificial General Intelligence). Αυτό συμβαίνει γιατί μοντελοποιείται η μάθηση που είναι ένα πολύ σημαντικό κομμάτι της διαδικασίας του αλγορίθμου.

Ένας σημαντικός τομέας των νευρωνικών δικτύων που έχει αναπτυχθεί τα τελευταία χρόνια λόγω της μεγάλης ζήτησης που υπάρχει σε τέτοιου είδους προβλήματα είναι ο τομέας της Βαθιάς Μηχανικής Μάθησης (Deep Learning). Αυτός ο κλάδος αποδίδει πολύ καλά σε προβλήματα είτε με πάρα πολλές μεταβλητές είτε σε τεράστιους όγκους δεδομένων είτε και στα δύο. Με αποτέλεσμα, να μπορούν σε πάρα πολλές εφαρμογές να δημιουργηθούν μοντέλα με πολύ καλή προβλεπτική ικανότητα.

Η μάθηση από ένα ή λίγα παραδείγματα είναι μία από τις βασικές ικανότητες του ανθρώπου από την πρώιμη βρεφική ηλικία, αλλά εξακολουθεί να αποτελεί σημαντικό πρόβλημα για αλγορίθμους των νευρωνικών δικτύων. Το πρόβλημα όμως που φαίνοταν να υπάρχει είναι, ότι είναι πολύ δύσκολα τέτοιου είδους μοντέλα να καταφέρουν να εξάγουν συμπεράσματα με πολύ λίγα δεδομένα. Σε αυτό το πρόβλημα μπορεί να δώσει απαντήσεις ένας νέος κλάδος των νευρωνικών δικτύων, η μάθηση με λίγα δεδομένα (Few shot learning).

Το θέμα της συγκεκριμένης διπλωματικής εργασίας είναι η χρήση νευρωνικών δικτύων που εξάγουν συμπεράσματα με τη χρήση πολύ λίγων δεδομένων. Τα δίκτυα στα οποία θα επικεντρωθούμε κάνουν χρήση των κανόνων της Μπεϋζιανής Στατιστικής. Οπότε, αρχικά, θα αναφέρουμε θεμελειώδης αρχές της Μπεϋζιανής Στατιστικής και στη συνέχεια θα αναφέρουμε επιγραμματικά διαφορετικές αρχιτεκτονικές για αλγορίθμους που χρησιμοποιούν λίγα δεδομένα. Στο κύριο κομμάτι της διπλωματικής, θα ασχοληθούμε με τη χρήση της Μπεϋζιανής Στατιστικής σε αυτούς τους αλγορίθμους και στην αναπαραγωγή αποτελεσμάτων με τη χρήση αντίστοιχων νευρωνικών δικτύων.

2 Μπεϋζιανή Στατιστική

2.1 Εισαγωγή

Η Μπεϋζιανή Στατιστική στηρίζεται σε ένα διαφορετικό πλαίσιο από την Κλασική Στατιστική. Στη Μπεϋζιανή Στατιστική, οι παράμετροι ενός προβλήματος θεωρούνται τυχαίες μεταβλητές. Αυτή η διαφορά είναι πολύ σημαντική και θέτει όλο το πλαίσιο της στατιστικής μελέτης σε διαφορετική βάση. Σε ένα κλασικό πρόβλημα στατιστικής για μια παράμετρο ϑ ενός πληθυσμού, η στατιστική ανάλυση θα βασιζόταν στη συνάρτηση $f(x|\vartheta)$, δηλαδή στη συνάρτηση της κατανομής πιθανότητας του δείγματος δεδομένης της παραμέτρου. Στη Μπεϋζιανή Στατιστική, η $f(\vartheta|x)$, η συνάρτηση κατανομής πιθανότητας της παραμέτρου δεδομένου του δείγματος, είναι το αντικείμενο της ανάλυσης, γεγονός που προκύπτει από το ότι η παράμετρος θα αντιμετωπίζεται ως τυχαία μεταβλητή. Σε πολλές περιπτώσεις αυτή η οπτική βοηθάει αρκετά στην απεικόνιση των δεδομένων του προβλήματος (2). Απαραίτητος στη Μπεϋζιανή Στατιστική είναι και ο ορισμός μια εκ των προτέρων (prior) κατανομής για την παράμετρο ϑ του δείγματος. Αυτή η εκ των προτέρων κατανομή απεικονίζει την πρώτερη γνώση που έχουμε για το πρόβλημα χωρίς να χρησιμοποιούνται τα δεδομένα του προβλήματος. Στην παρακάτω παράγραφο θα αναφέρουμε, λοιπόν, μερικά χαρακτηριστικά όπως η εκ των προτέρων κατανομή κατανομή, που δείχνουν τις διαφορές της Μπεϋζιανής Στατιστικής με την Κλασική Στατιστική.

Στη Μπεϋζιανή Στατιστική υποστηρίζεται η έλλειψη αντικειμενικότητας των αποτελεσμάτων και αν υπάρχει κάποια εκ των προτέρων πληροφορία για την παράμετρο, πρέπει να εισάγεται στη Στατιστική Ανάλυση. Στην Κλασική Στατιστική αντιθέτως δεν συμμερίζονται την παραπάνω άποψη και η θεωρία βασίζεται στον ορισμό των πιθανοτήτων σύμφωνα με τους κλασικούς ορισμούς που υπακούν στο Νόμο των Μεγάλων Αριθμών. Οπότε δύο αρχικές διαφορές είναι η Αρχική Πληροφορία και η Υποκειμενική Πιθανότητα (2). Αρχικά, όπως αναφέραμε και στην προηγούμενη παράγραφο η Αρχική Πληροφορία είναι από τις βάσεις της Μπεϋζιανής Στατιστικής και δίνει το δικαίωμα να αντιμετωπίζεται το κάθε πρόβλημα διαφορετικά κάτι το οποίο δεν συνηθίζεται στην Κλασική Στατιστική ειδικά όταν μιλάμε για προβλήματα που φαινομενικά έχουν τα ίδια χαρακτηριστικά. Με παρόμοιο τρόπο συμβαίνει το ίδιο και στον ορισμό της Υποκειμενικής Πιθανότητας. Οι νόμοι που ακολουθεί η Κλασική Στατιστική οδηγούν πολύ συχνά σε πολύπλοκα συμπεράσματα. Οπότε δεν δίνεται η ευελιξία στο χρήστη να παράγει κάτι υποκειμενικό. Στη Μπεϋζιανή Στατιστική αντίθετα η ύπαρξη τηςεκ των προτέρων κατανομής που είναι και η βάση των συμπερασμάτων της Στατιστικής Ανάλυσης δίνει τη δυνατότητα της υποκειμενικής μελέτης ενός προβλήματος.

Η επόμενη μεγάλη διαφορά απορρέει από τον ορισμό της παραμέτρου ϑ από τη μπεϋζιανή Στατιστική ως τυχαία μεταβλητή. Ο ορισμός αυτός οδηγεί τα συμπεράσματα που βγαίνουν από μια τέτοια ανάλυση να πηγάζουν από τη θεωρία πιθανοτήτων. Αυτό

το γεγονός συμβαίνει γιατί όλα τα αποτελέσματα μπορούν να γραφτούν σαν εκφράσεις της παραμέτρου ϑ, οι οποίες είναι συμπεράσματα της εκ των υστέρων κατανομής. Στην Κλασική Στατιστική γίνεται η κριτική για την παρουσία αυθαίρετων κριτιριών απόφασης (2). Τα κριτήρια αυτά είναι γνωστά και πολλές φορές χρησιμοποιούνται για τον προσδιορισμό εκτημητριών ενός προβλήματος. Σε πολλές περιπτώσεις λοιπόν σε μία ανάλυση της Κλασικής Στατιστικής αναφέρονται σε μία καλή ή κακή εκτιμήτρια για μία παράμετρο ενός προβλήματος, από ένα αυθαίρετο κριτήριο απόφασης . Η Μπεϋζιανή Στατιστική δεν χρησιμοποιεί αυθαίρετα κριτήρια για την αξιολόγηση και σύγκριση εκτιμητριών, αντίθετα στηρίζεται στην εκ των υστέρων κατανομή για να εξάγει συμπεράσματα για την τυχαία μεταβλητή ϑ. Η διαφορά λοιπόν των δύο προσεγγίσεων έγκειται στο πώς και με ποιο τρόπο θα εκτιμηθεί η παράμετρος ϑ ενός προβλήματος για να βγούν τα στατιστικά συμπεράσματα, είτε στη μία περίπτωση σαν τυχαία μεταβλητή (Μπεϋζιανή Στατιστική), είτε με μια εκτιμήτρια (Κλασική Στατιστική).

Όπως αναφέραμε νωρίτερα, η Μπεϋζιανή Στατιστική χειρίζεται την παράμετρο ϑ ως τυχαία μεταβλητή σε αυτή τη διαδικάσια υπάρχουν τέσσερα χαρακτηριστικά βήματα που θα παραθέσουμε:

1. Καθορισμός της εκ των προτέρων κατανομής $f(\vartheta)$,
2. Καθορισμός του μοντέλου πιθανοφάνειας $f(x \mid \vartheta)$,
3. Υπολογισμός της εκ των υστέρων κατανομής $f(\vartheta \mid x)$ από το θεώρημα του Bayes,
4. Εξαγωγή συμπερασμάτων από την εκ των υστέρων κατανομή.

Στην συνέχεια, παραθέτεται το θεώρημα του Bayes στο οποίο παρουσιάζεται, ο τρόπος με τον οποίο συνδυάζονται τα δεδομένα με τις εκ των προτέρων πεποιθήσεις του αναλυτή προκειμένου να παραχθεί η εκ των υστέρων κατανομή στην οποία εμφωλευέται όλη η πληροφορία για την άγνωστη παράμετρο. Επιπλέον, μιας και η επιλογή των εκ των προτέρων κατανομών καμιά φορά είναι πολύ δύσκολη ή και αδύνατη. Η διατύπωση του θεωρήματος Bayes θα γίνει σε μία μορφή κατάλληλη για τυχαίες μεταβλητές αντί για ενδεχόμενα και θα αναφέρουμε ορισμένα προβήματα που προκύπτουν όταν προσπαθούμε να χρησιμοποιήσουμε αυτό το αποτέλεσμα στο πλαίσιο της συμπερασματολογίας για μία παράμετρο ϑ (2).

2.2 Το Θεώρημα του Bayes

Το θεώρημα του Bayes, στο αποτέλεσμα του οποίου βασίζεται μια τεχνική που θα παρουσιαστεί με τη χρήση της άγνωστης παραμέτρου ϑ, διατυπώνεται σε όρους τυχαίων μεταβλητών με συναρτήσεις πυκνότητας πιθανότητας f και παίρνει τη μορφή:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)}$$

όπου:

$$x^T = (x_1, x_2, ..., x_n) \text{ το τυχαίο δείγμα.}$$

$$\theta = (\theta_1, \theta_2, ..., \theta_m) \in \Theta \text{ το διάνυσμα των άγνωστων παραμέτρων.}$$

.

$f(x|\theta)$ η συνάρτηση πιθανοφάνειας, η οποία εκφράζει την πιθανότητα παρατήρησης διαφορετικών x_i κάτω από διαφορετικές τιμές της άγνωστης παραμέτρου ϑ, με $f(x_i|\theta)$ να είναι η συνάρτηση πυκνότητας ή μάζας πιθανότητας που περιγράφει την τυχαία μεταβλητή x_i.

$f(\theta)$ η εκ των προτέρων κατανομή για την άγνωστη παράμετρο, η οποία βασίζεται σε πληροφορίες που υπάρχουν από προηγούμενες έρευνες για την παράμετρο ϑ, ή σε πεποιθήσεις που έχουμε γι αυτή, τη δεδομένη χρονική στιγμή που μελετάτε το πρόβλημα.

$f(\theta|x)$ η εκ των υστέρων κατανομή που περικλείει όλη τη στατιστική συμπερασματολογία για την αγνωστη παράμετρο ϑ και προκύπτει ως το αποτέλεσμα αναπροσαρμογής των δεδομένων με την εκ των προτέρων γνώση.

Σημειώνεται ότι στην περίπτωση που η άγνωστη παράμετρος ϑ είναι διακριτή, το ολοκλήρωμα στον παρονομαστή αντικαθίσταται από το άθροισμα:

$$\sum_i f(\theta_i)f(x|\theta_i)$$

2.3 Εναλλακτική Μορφή του Θεωρήματος του Bayes

Είναι εύκολο να παρατηρηθεί ότι ο παρανομαστής στο θεώρημα του Bayes είναι συνάρτηση μόνο του x, κατά συνέπεια δεν εξαρτάται καθόλου από το ϑ που φεύγει από το ολοκ-

λήρωμα. Συνεπώς ένας ενναλακτικός τρόπος γραφής του Θεωρήματος Bayes είναι ο παρακάτω:

$$\begin{aligned} f(\theta|x) &= cf(\theta)f(x|\theta) \\ &f(\theta)f(x|\theta) \\ &= h(\theta) \end{aligned}$$

Αυτή η σχέση, δείχνει οτι η εκ των υστέρων κατανομή είναι ανάλογη με την εκ των προτέρων κατανομή πολλαπλασιαζόμενη με την πιθανοφάνεια. Ακόμα η σταθερά c που εξαρτάται μόνο από το x (όχι από το ϑ), είναι μια σταθερά κανονικοποίησης, που είναι ορισμένη έτσι ώστε η εκ των υστέρων κατανομή να ολοκληρώνει στη μονάδα

Υπάρχει μοναδική σ.π.π., έστω $g(\theta)$, που είναι ανάλογη σε κάποια δοσμένη συνάρτηση $h(\theta)$ για $\theta \in \Theta$, αφού η $g(\theta)$ μπορεί να καθοριστεί μονοσήμαντα ως $g(\theta) = ch(\theta)$, όπου:

$$c = \frac{1}{\int_\Theta h(\theta)d\theta}.$$

2.4 Επιλογή Κατάλληλου Μοντέλου Πιθανοφάνειας

Το παραπάνω αποτέλεσμα θα χρησιμοποιηθεί αρκετές φορές στους αλγορίθμους μας για να αναγνωρίστουν γνωστές σ.π.π. που είναι ανάλογες σε συγκεκριμένες εκ των υστέρων κατανομές. Ακόμα λόγω αυτής της σχέσης υπάρχει η δυνατότητα να αφαιρεθεί όποιος παράγοντα της συνάρτησης $g(\theta) = ch(\theta)$ δεν εξαρτάται από το ϑ πριν από τη διαδικασία της κανονικοποίησης.

Ένα ακόμα ζήτημα είναι η επιλογή του κατάλληλου μοντέλου πιθανοφάνειας. Η επιλογή αυτή εξαρτάται από το εκάστοτε πρόβλημα που πρέπει να αντιμετωπιστεί. Είναι ουσιαστικά το ίδιο πρόβλημα που υπάρχει στην στατιστική όταν δεν είναι γνωστό ποιο μοντέλο πρέπει να επιλεχτεί, είτε στο χώρο της Μηχανικής Μάθησης, ποιός αλγόριθμος ταιριάζει περισσότερο στο εκάστοτε πρόβλημα. Ένας παράγοντας που μπορεί να δείξει πιο είναι το κατάλληλο μοντέλο για να χρησιμοποιηθεί είναι ο τρόπος επιλογής των δεδομένων. Ένας άλλος είναι τα χαρακτηριστικά που έχει μια επεξηγηματική μεταβλητή και τι κατανομές μπορεί να ακολουθεί. Σαφώς και η κατανομή που ακολουθεί η ίδια η επεξηγηματική μεταβλητή, μπορεί να είναι είτε συνεχής είτε διακριτή. Ακόμα και αν είναι γνωστό αν είναι συνεχής, τότε το μοντέλο μπορεί να αλλάζει πάλι ανάλογα με την κατανομή που γνωρίζουμε οτι ακολουθεί το πρόβλημα (1).

Η εφαρμογή του θεωρήματος του Bayes μπορεί να θεωρηθεί απλή αν και υπολογιστικά δύσκολη λόγω της σταθεράς κανονικοποίησης c. Βέβαια με κατάλληλη επιλογή

των εκ των προτέρων κατανομών και της πιθανοφάνειας μπορεί ο υπολογισμός του συγκεκριμένου ολοκληρώματος να αποφευχθεί.

Στη μπεϋζιανή ανάλυση δίνεται ένας πιο πλήρες συμπέρασμα σε σχέση με την Κλασική Στατιστική, αφού όλη η πληροφορία για το θ, η οποία είναι διαθέσιμη από την εκ των προτέρων κατανομή και τα δεδομένα, απεικονίζεται στην εκ των υστέρων κατανομή. Αυτό έχει ως συνέπεια η $f(\theta|x)$ να περιέχει όλο το συμπέρασμα για το εκάστοτε πρόβλημα. Βέβαια, συχνά αυτό το συμπέρασμα απεικονίζεται υπο τη μορφή μιας σημειακής εκτίμησης η ενος διαστήματος εκτίμησης (5). Επιπλέον, κάποιες από τις επιθυμητές ιδιότητες των στατιστικών συναρτήσεων, δηλαδή των συναρτήσεων των δεδομένων που χρησιμοποιούνται για τους σκοπούς της συμπερασματολογίας, υπάρχουν και στη Μπεϋζιανή Ανάλυση. Για παράδειγμα στην Κλασική Στατιστική γίνεται συχνά η συζήτηση για την ύπαρξη επαρκών συναρτήσεων για την αντιμετώπιση όλων των πιθανών προβλημάτων. Αυτή η συζήτηση με παρόμοιο τρόπο μπορεί να γίνει και για τη μπεϋζιανή συμπερασματολογία. Για παράδειγμα, στη Μπεϋζιανή Στατιστική, η επάρκεια μπορεί να χαρακτηριστεί λέγοντας ότι αν πάρουμε μία συνάρτηση $h(x)$ των δεδομένων μας, τότε η $h(x)$ είναι επαρκής για το θ αν η εκ των υστέρων κατανομή $f(\theta|x)$ εξαρτάται από τα δεδομένα μόνο μέσω της $h(x)$ και όχι από τις επιμέρους τιμές των x_i (3).

2.5 Επιλογή Κατάλληλης Εκ των Προτέρων Κατανομής

Ένα θεμελειώδες ζήτημα στη Μπεϋζιανή Στατιστική είναι κατάλληλη επιλογή της εκ των προτέρων κατανομής θα παρατεθούν στη συνέχεια μερικά σημεία που πρέπει να προσεχθούν για την επιλογή μιας εκ των προτέρων κατανομής και στη συνέχεια θα γίνει εκτενή αναφορά για τον τρόπο επιλογής (2).

Η εκ των προτέρων κατανομή είναι η κατανομή που έχει η παράμετρος θ πριν την εκτίμηση του προβλήματος. Αυτή η κατανομή προσδιορίζεται από πολλούς διαφορετικούς παράγοντες. Όμως, το πιο βασικό χαρακτηριστικό της Μπεϋζιανής οπτικής είναι ότι αυτή η κατανομή μπορεί να διαφέρει από αναλυτή σε αναλυτή αφού εμπεριέχει την πρώτερη γνώση του για το εκάστοτε πρόβλημα. Το αποτέλεσμα όμως για την παράμετρο θ που εξετάζεται εξαρτάται κυρίως από τα δεδομένα που θα εισαχθούν στο πρόβλημα. Άρα η χρήση μιας κατανομής που ίσως να μην ταιριάζει απόλυτα στην παράμετρο θ δεν αποτελεί πρόβλημα στον υπολογισμό της. Στη μόνη περίπτωση που θα αντιμετωπίστει πρόβλημα είναι στη χρήση μιάς εντελώς παράλογης κατανομής για την παράμετρο θ .

Στις περισσότερες περιπτώσεις, ο αναλυτής έχει μία πρώτερη γνώση για το πρόβλημα που θέλει να μελετήσει. Οπότε θα χρησιμοποιήσει κάποια εκ των προτέρων κατανομή που, δεν θα δημιουργήσει πρόβλημα στον υπολογισμό της παραμέτρου θ και θα είναι σε μια πιο απλή μορφή για να γίνουν πιο απλοί οι μαθηματικοί υπολογισμοί. Σε περίπτωση,

που δεν γνωρίζει ο αναλυτής αρκετές πληροφορίες για την εκ των προτέρων κατανομή μπορεί να δηλώσει την άγνοια του μέσω της επιλογής του. Για παράδειγμα θα μπορούσε να επιλέξει μία ομοιόμορφη κατανομή. Με αυτόν τον τρόπο θα γνωρίζει ότι δεν επέλεξε μια εντελώς παράλογη κατανομή για την παράμετρο θ .

Οι εκ των προτέρων (prior) κατανομές εκφράζουν τη γνώση που υπάρχει για την άγνωστη παράμετρο θ , πριν από τη συλλογή των δεδομένων (4). Αυτό το γεγονός έχει σαν αποτέλεσμα, κάθε τέτοια ανάλυση να μπορεί να χαρακτηριστεί ώς υποκειμενικη εάν η επιλογή της εκ των προτέρων κατανομής δεν έχει γίνει με σωστό τρόπο. Αν και διαφορετικές εκ των προτέρων κατανομές οδηγούν σε διαφορετικά αποτελέσματα με κατάλληλη επιλογή δεν φαίνεται η επίδραση τους στο εκάστοτε πρόβλημα. Μια λογική επιλογή εκ των προτέρων κατανομών, γίνεται με τέτοιο τρόπο ώστε αυτές να πληρούν τις ακόλουθες δύο προϋποθέσεις. Πρώτον, να χρησιμοποιούν την πληροφορία από διάφορες έρευνες πού έχουν γίνει στο συγκεκριμένο γνωστικό αντικείμενο και τη γνώμη των ειδικών. Δεύτερον, να χρησιμοποιείται κάποια από τις γνωστές οικογένειες κατανομών για να διευκολύνθουν οι υπολογισμοί για το Θεώρημα του Bayes (2).

Όπως αναφέρθηκε και προηγουμένως υπάρχουν περιπτώσσεις όπου σε ένα πρόβλημα που είναι προς μελέτη δεν υπάρχει πρότερη πληροφορία είτε ακόμα δεν είναι εύκολο να χρησιμοποιηθεί η γνώμη των ειδικών. Σε αυτές τις περιπτώσεις προτιμάτε να χρησιμοποιείται κάποια κατανομή που να είναι μη πληροφοριακή και να υπόκειται κυρίως στα δεδομένα για τον υπολογισμό της εκ των υστέρων κατανομής. Όποτε, επιλέγονται κυρίως κατανομές που να έχουν μεγάλη διασπορά (2).

Μερικές διαδεδομένες κατηγορίες εκ των προτέρων κατανομών που είναι σύνηθες να χρησιμοποιούνται είναι οι εξής:

1. Οι συζυγείς εκ των προτέρων κατανομές
2. Η μη πληροφοριακή εκ των προτέρων κατανομή του Jeffrey
3. Οι ιεραρχικές εκ των προτέρων κατανομές
4. Οι εκ των προτέρων κατανομές που βασίζονται σε δυνάμεις της πιθανοφάνειας

3 Μάθηση με λίγα δεδομένα

3.1 Εισαγωγή

Ένας άνθρωπος για να αναγνωρίσει μια καινούργια κλάση αντικειμένων χρειάζεται πολύ λίγα παραδείγματα. Σε αντίθεση, οι περισσότεροι αλγόριθμοι μηχανικής μάθησης βασίζονται στην ύπαρξη πολλών δεδομένων, που μπορούν να τους βοηθήσουν στην εξαγωγή μοτίβων για να προσδιορίσουν καλύτερα μία κλάση αντικειμένων. Ο στόχος της μάθησης με λίγα δεδομένα είναι η εξαγωγή συμπερασμάτων και η σωστή ταξινόμηση αντικειμένων από πολύ λίγα δεδομένα για την κάθε κλάση, σε κάποιες περιπτώσεις και μόνο μία παρατήρηση για την κάθε κλάση (one-shot-learning). Οπότε η ανάγκη για τη δημιουργία τέτοιων αλγορίθμων, προέρχεται από περιπτώσεις προβλημάτων που είναι είτε πολύ δύσκολη η εύρεση πολλών δεδομένων ή πολύ ακριβή η παραγωγή τους (7). Υπάρχουν πολλές κατηγορίες τέτοιων προβλημάτων που έχουν αυτήν την ανάγκη όπως η κατηγοριοποίηση εικόνων, η εύρεση χαρακτήρων από ένα αλφάβητο είτε ακόμα και σε περιπτώσεις εξαγωγής φαρμάκων σύμφωνα με (7). Βέβαια, επειδή σαν διαδικασία είναι αρκετά συμφέρουσα μπορεί να αποδειχτεί και χρήσιμη σε προβλήματα που δεν έχει ακόμα χρησιμοποιηθεί αφού με τη χρήση λίγων δεδομένων μπορεί να εξάγει ασφαλή συμπεράσματα (7).

3.2 Ορισμός ενός προβλήματος μάθησης με λίγα δεδομένα

Θα παρατεθεί ένας ορισμός όχι για τους αλγορίθμους που χρησιμοποιούνται σε αυτό το χώρο αλλά για τη δομή των προβλημάτων που καλούνται να αντιμετωπίσουν αυτοί οι αλγόριθμοι.

Ένα πρόβλημα με το οποίο ασχολείται ο τομέας της μάθησης με λίγα παραδείγματα (Few Shot Learning) θα έχει λίγα δεδομένα. Αυτά τα δεδομένα θα απεικονίζουν διαφορετικά αντικείμενα που θα ορίζονται μέσα από τις κλάσεις του προβλήματος. Τον αριθμό των διαφορετικών κλάσεων θα το συμβολίζουμε με n. Ο αριθμός των κλάσεων ενός τέτοιου προβλήματος μπορεί να είναι πολύ μεγάλος (το λιγότερο 2 κλάσεις). Ας υποθέσουμε αρχικά ότι όλες οι κλάσεις θα υπάρχουν τις ίδιες φορές. Αυτό, βέβαια, δεν ισχύει σε όλες τις περιπτώσεις. Θα συμβολίζουμε με k, τον αριθμό των φορών που εμφανίζεται η κάθε κλάση, στα δεδομένα που θα εκπαιδεύσουμε.

Στο πρόβλημα θα υπάρχουν τρία σύνολα δεδομένων:

1. Το σύνολο των δεδομένων που θα εκπαιδευτεί.

2. Το σύνολο των δεδομένων που θα ελεγχθεί το μοντέλο.

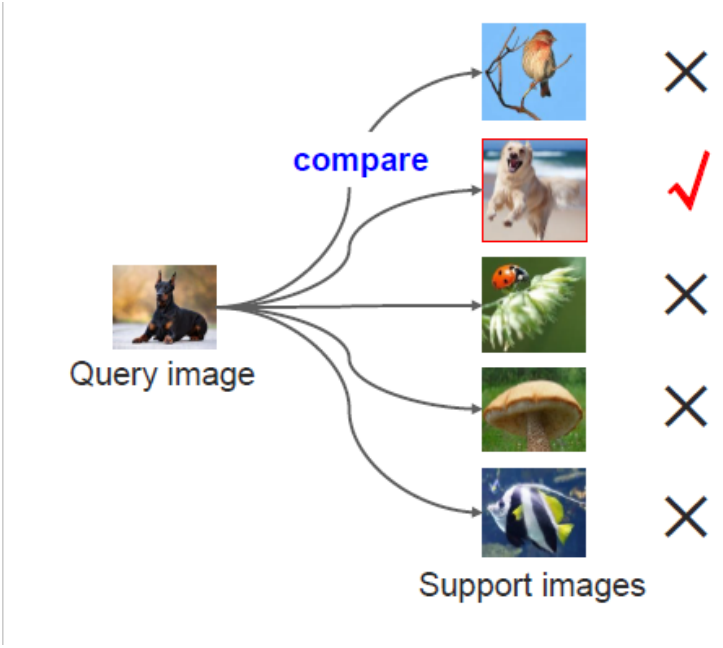
3. Ένα τρίτο σύνολο που περιέχει κατηγορίες που δεν υπάρχουν στα δεδομένα που θα εκπαιδευτούν.

Αν υπάρχουν k διαφορετικές παρατηρήσεις για την κάθε μία από τις n διαφορετικές κλάσεις σε ένα σύνολο δεδομένων, τότε αυτό θα ονομάζεται N-way k-shot πρόβλημα.

3.3 Μέθοδοι Μάθησης βασισμένες στη μετρική

Σε όλες τις εφαρμογές της Μηχανικής Μάθησης είναι απαραίτητη η χρήση μίας μετρικής για να μπορεί να οριστεί η απόσταση μεταξύ των παρατηρήσεων του προβλήματος. Αυτή η μετρική μπορεί να είναι μια από τις γνωστές μετρικές απόστασεις (π.χ. Ευκλείδια μετρική). Όμως, υπάρχουν περιπτώσεις που είναι δύσκολο να βρεθεί μια κατάλληλη μετρική για να αποσαφηνίσει την απόσταση των παρατηρήσεων (8).

Στις μεθόδους που βασίζονται στο μετρική-μάθηση (metric-learning), ο πρώτος στόχος είναι η κατασκευή τέτοιων μετρικών ακόμα και στις περιπτώσεις που υπάρχουν λίγα δεδομένα (8). Αυτό το σκοπό όμως, οι μέθοδοι προσπαθούν να το πετύχουν με τον εξής τρόπο: αρχικά, οι εικόνες ή οι παρατηρήσεις που υπάρχουν από κάθε κλάση να έχουν πολλά κοινά χαρακτηριστικά έτσι ώστε να δίνεται η δυνατότητα στον αλγόριθμο να τις ταξινομεί κατάλληλα. Άλλα για να πετύχει αυτό θα πρέπει να διαφέρουν και αρκετά από τις εικόνες των άλλων κλάσεων έτσι ώστε αυτά τα κοινά χαρακτηριστικά που αποκτούν να μην βρίσκονται ή τουλάχιστον να μη μοιάζουν αρκετά με τα χαρακτηριστικά οποιασδήποτε άλλης κλάσης.



Σχήμα 1: Σύγκριση των εικόνων

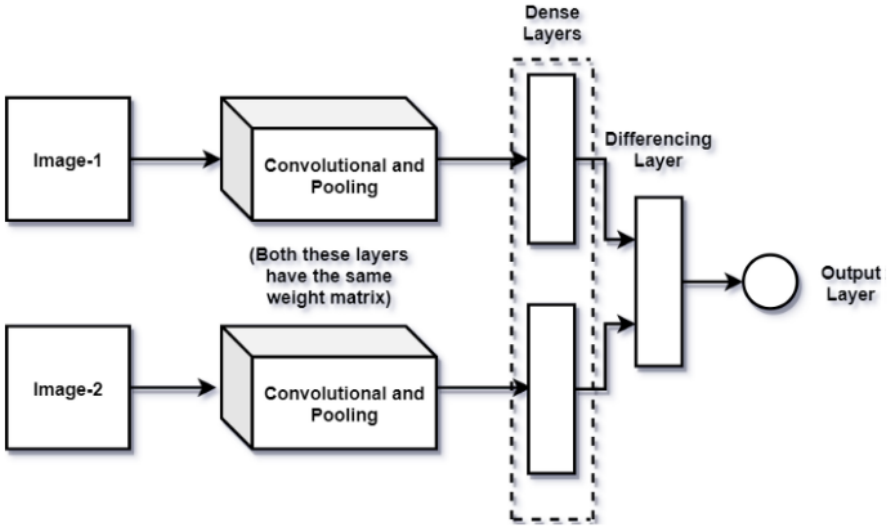
3.3.1 Σιαμαία Νευρωνικά Δίκτυα

Τα Σιαμαία Νευρωνικά Δίκτυα είναι μια κλάση νευρωνικών δικτύων που συναντάται σε εφαρμογές της μάθησης με λίγα δεδομένα (Few Shot Learning) και ειδικότερα είναι κατάλληλα για μάθηση με ένα δεδομένο ανά κλάση (One Shot Learning) (8).

Αρχικά ένα τέτοιο νευρωνικό δίκτυο παίρνει δύο διαφορετικές παρατηρήσεις ή εικόνες και δίνει σαν αποτέλεσμα το αν αυτές οι δύο εικόνες ανήκουν στην ίδια κλάση. Αυτή η σύγκριση γίνεται περνώντας συγχρόνως τις δύο εικόνες από το κάθε επίπεδο του δικτύου. Σαν μετρική χρησιμοποιείται η απόσταση που δημιουργείται από τα χαρακτηριστικά της κάθε εικόνας, καταλήγοντας στο αποτέλεσμα αν ανήκουν ή όχι στην ίδια κλάση. Κατά τη διάρκεια της εκμάθησης διαλέγονται οι κλάσεις για το κάθε παράδειγμα από ένα υπερσύνολο κλάσεων έτσι ώστε το δίκτυο να γενικεύει μεταξύ πολλών διαφορετικών περιπτώσεων και να μπορεί να ξεχωρίσει καλύτερα όλους τους πιθανούς συνδυασμούς μη κοινών κλάσεων. Για αυτό το σκοπό κατά τη διάρκεια της εκμάθησης χρησιμοποιούνται κλάσεις που διαφέρουν αρκετά μεταξύ τους.

Ακόμα, τα Σιαμαία Νευρωνικά Δίκτυα έχουν παρόμοια αρχιτεκτονική με τα συνελικτικά και τα συγκεντρωτικά δίκτυα εκτός από το ότι δεν έχουν ένα softmax επίπεδο. Έτσι σταματάνε στα πυκνά (dense) επίπεδα (8). Όπως εξηγήσαμε και στην αρχή αφού το δίκτυο ξεκινάει με είσοδο δύο εικόνες είναι και λογικό επακόλουθο σαν έξοδο να έχει δύο πυκνά επίπεδα. Στο τέλος, λοιπόν, υπολογίζει τη διαφορά μεταξύ αυτών των δύο επιπέδων και δίνει σαν αποτέλεσμα σε απλό νευρωνικό δίκτυο την απάντηση μέσω μίας σιγμοειδούς συνάρτησης. Για αυτόν το λόγο το αντίστοιχο σύνολο εκπαίδευσης των δεδομένων σε τέτοιου είδους προβλήματα πρέπει να αποτελείται κάθε φορά από παρατηρήσεις που θα έχουν δύο εικόνες και μια τιμή που θα συμβολίζει με 0 ή 1 αν αυτές οι δύο εικόνες ανήκουν στην ίδια κλάση.

Ακριβώς με ένα τέτοιο τρόπο έγινε και η παραγωγή ενός τέτοιου δικτύου στην εργασία (8). Στη συγκεκριμένη δημοσίευση, οι Koch, Zemel και Salakhutdinov εκπαιδεύσαν ένα μοντέλο το οποίο δέχεται στην είδοδό του δύο είκονες και δίνει σαν αποτέλεσμα την πιθανότητα να ανήκουν στην ίδια κλάση. Αυτές οι δύο εικόνες περνούν συγχρόνως από ένα δίκτυο και η κάθε μια ανάλογα με το αποτέλεσμα της κατατάσσεται σε μια κλάση του προβλήματος. Στη συνέχεια, υπολογίζεται η διαφορά μεταξύ αυτών των δύο αποτελεσμάτων σε έναν αριθμό. Αυτός ο αριθμός περνάει μετά απο μία σιγμοειδή συνάρτηση που τον κατατάσσει σε 0 ή 1 με τη χρήση του λάθους της εντροπίας της πληροφορίας (cross-entropy loss). Το τελευταίο βήμα γίνεται για να καταλήξει το δίκτυο αν οι δύο εικόνες ανήκουν στην ίδια κλάση.



Σχήμα 2: Σιαμαία Νευρωνικά Δίκτυα

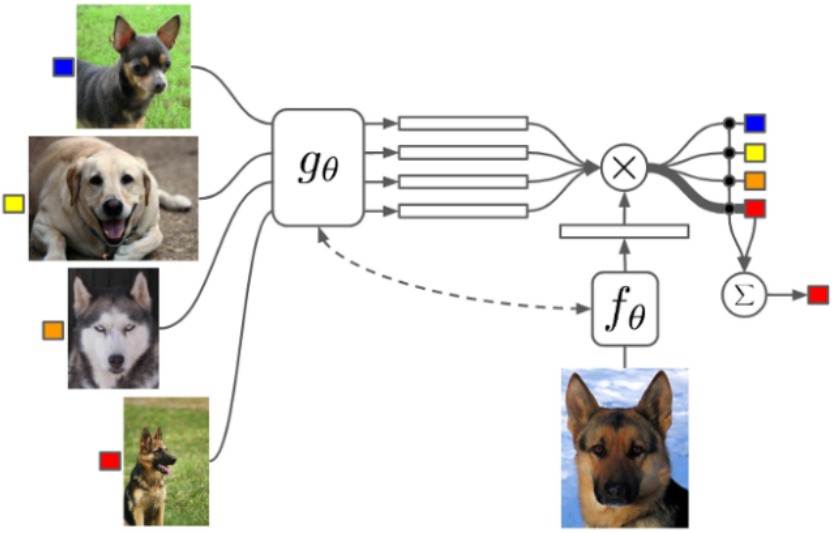
3.3.2 Δίκτυα Ταιριάσματος

Τα Δίκτυα Ταιριάσματος αποτελούν την πρώτη προσπάθεια που έγινε στα νευρωνικά δίκτυα (για λίγα δεδομένα), έτσι ώστε να υπάρχουν στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου του μοντέλου, η δομή των n-κλάσεων k-παραδειγμάτων που είναι η βάση της μάθησης με λίγα δεδομένα. Η λογική είναι πολύ απλή και ταιριάζει αρκετά με λογικές που υπάρχουν στους περισσότερους αλγορίθμους μηχανικής μάθησης. Το σύνολο εκπαίδευσης αποτελείται από διαφορετικά δεδομένα της κάθε κλάσης και προσπαθεί να βρεί κάποια κοινά χαρακτηριστικά για την κάθε κλάση έτσι ώστε να ταιρίαζει τα αντικείμενα που ανήκουν στην ίδια κατηγορία.

Στην περίπτωση των Σιαμαίων Νευρωνικών Δικτύων είδαμε ότι γινόταν η ταξινόμηση του αν ανήκουν δύο εικόνες σε μία κλάση σε δύο στάδια. Πρώτα, βλέπαμε κατά πόσο είναι κοντά μεταξύ τους και μετά σε δεύτερο επίπεδο εάν αυτή η απόσταση που έχουν αποκτήσει τις καθιστά ικανές να ανήκουν ή όχι στην ίδια κλάση. Αυτή η διαδικασία όμως δεν μπορεί να είναι βέλτιστη γιατί η συνάρτηση που καθορίζει την απόσταση που θα έχουν μεταξύ τους οι δύο εικόνες μετριέται από μία άλλη μετρική σε επόμενο επίπεδο. Σε αντίθεση τα Δίκτυα Ταιριάσματος χρησιμοποιούν τον (διαφορικό ταξινομητή πλησιέστερου γείτονα) για να δημιουργείται και η μετρική της απόστασης αλλά και το αν ανήκει μια εικόνα σε μια συγκεκριμένη κλάση στο ίδιο βήμα (9).

Ένα παράδειγμα χρήσης του συγκεκριμένου δικτύου προτάθηκε στην εργασία (9). Στο συγκεκριμένο πρόβλημα, οι (9) προσπάθησαν να εκπαιδεύσουν Δίκτυα Ταιριάσματος τα οποία καλούνταν στη συνέχεια να ταξινομήσουν n αντικείμενα από k κλάσεις που

δεν προϋπήρχαν στο σύνολο εκπαίδευσης. Βέβαια η δυνατότητα των Δικτύων Ταιριάσματος να βρίσκουν τις ομοιότητες μεταξύ των παρατηρήσεων του συνόλου υποστήριξης είτε των δεδομένων από το σύνολο εκπαίδευσης δεν καλύπτουν την αδυναμία των δικτύων στην περίπτωση που στο πρόβλημα που είναι πρός αντιμετώπιση υπάρχουν μη ισορροπημένα (inbalanced) δεδομένα. Γιατί στην περίπτωση που στο σύνολο υποστήριξης δεδομένων υπάρχουν περισσότερα δεδομένα από μια συγκεκριμένη κλάση τότε το πιο πιθανό είναι να ταξινομήσει όλα τα δεδομένα στη συγκεκριμένη κλάση που υπερτερεί (9).



Σχήμα 3: Δίκτυα Ταιριάσματος

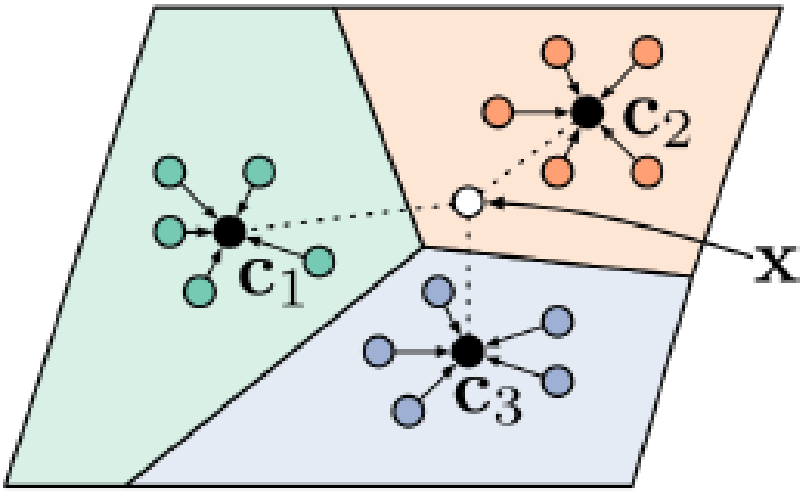
3.3.3 Πρωτότυπα Δίκτυα

Μία άλλη αρχιτεκτονική που χρησιμοποιείται ιδιαίτερα σε εφαρμογές της μάθησης με λίγα δεδομένα είναι τα Πρωτότυπα Δίκτυα. Τα συγκεκριμένα δίκτυα είναι ικανά να αντιμετωπίσουν το πρόβλημα των μη ισορροπημένων δεδομένων σε αντίθεση με τα Δίκτυα Ταιριάσματος, λόγω της κατασκευής τους. Η κατασκευή τους βασίζεται στην ύπαρξη ενός πρωτοτύπου για την κάθε κλάση το οποίο μπορεί να δημιουργηθεί ακόμα και αν η συγκεκριμένη κλάση έχει μόνο μια παρατήρηση, οπότε η δημιουργία του πρωτοτύπου είναι ανεξάρτητη από τον αριθμό των δεδομένων της κάθε κλάσης.

Ουσιαστικά, δημιουργούν για την κάθε κλάση μία μέση εικόνα η οποία παράγεται από το μέσο όρο των παρατηρήσεων της κάθε κλάσης που βρίσκεται στο σύνολο των δεδομένων εκπαίδευσης (10). Αρχικά, λοιπόν για την κάθε κλάση δημιουργείται ένα πρωτότυπο της κλάσης που αποτελεί την παρατήρηση που την εκπροσωπεί. Αυτή η

διαδικασία είναι η βάση για τη μετέπειτα ταξινόμηση των παρατηρήσεων σε κλάσεις. Η ομοιότητα τώρα κάθε καινούργιας παρατήρησης ισούται με την αρνητική έκφραση της ευκλείδιας νόρμας (10). Αυτό έχει ως αποτέλεσμα όσο μεγαλύτερος είναι ο αριθμός τόσο πιο πιθανό είναι να ανήκεις στη συγκεκριμένη κλάση. Αυτές οι αποστάσεις περνάνε μέσα από μια softmax συνάρτηση για να δώσουν την πιθανότητα να ανήκει μία εικόνα σε μία κλάση.

Συνεπώς ένα τέτοιο μοντέλο μαθαίνει καλά ένα μετρικό χώρο, όπου η μέση τιμή των παρατηρήσεων μιας κλάσης αποτελεί καλή αναπαράσταση για την κλάση και το ,αν ανήκει μια καινούργια παρατήρηση σε αυτή την κλάση μπορεί να προσδιοριστεί από την απόσταση της από το κέντρο της κλάσης. Σε αυτά τα νευρωνικά δίκτυα ,όπως γίνεται εύκολα αντιληπτό είναι κομβικής σημασίας η επιλογή της μετρικής που στις περισσότερες περιπτώσης είναι η ευκλείδια. Ακόμα σε ένα τέτοιο δίκτυο όσο περισσότερες κλάσεις υπάρχουν στο σύνολο εκπαίδευσης τόσο καλύτερα αποτελέσματα θα έχει αφού θα έχει περισσότερους μέσους όρους για να ταξινομήσει μια καινούργια παρατήρηση. Λειτουργεί καλύτερα στις περιπτώσεις που το σύνολο εκπαίδευσης και το σύνολο που θα γίνει ό έλεγχος έχουν δεδομένα από τις ίδιες κλάσεις (10). Στη συγκεκριμένη δημοσίευση (11) χρησιμοποίησαν τα Πρωτότυπα Δίκτυα σε εφαρμογή με μη ισορροπημένα δεδομένα τα οποία προέρχονταν είτε από κατηγορίες που υπήρχαν στο σύνολο ελέγχου ή ακόμα και από κατηγορίες που δεν υπήρχαν καν σε αυτό.

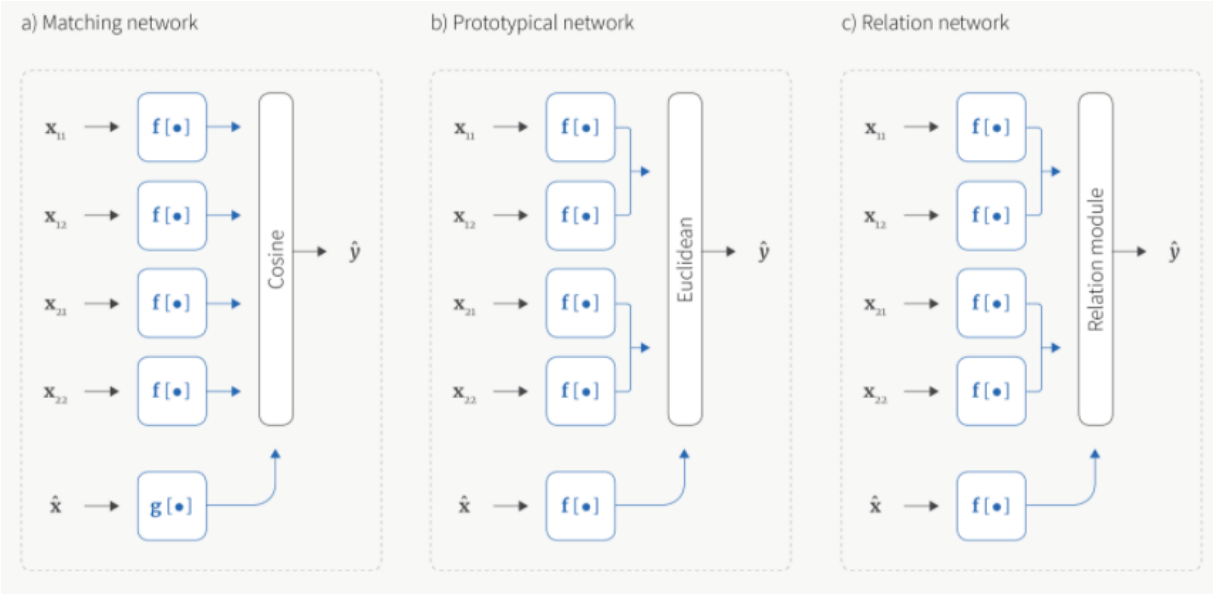


Σχήμα 4: Πρωτότυπα Δίκτυα

3.3.4 Σχεσιακά Δίκτυα

Οι δυο προηγούμενες αρχιτεκτονικές νευρωνικών δικτύων χρησιμοποιούν μια προκαθορισμένη μετρική όπως την Ευκλείδια για να ταξινομήσουν τις διάφορες κατηγορίες. Τα Σχεσιακά Δίκτυα (Relation Network) κατά τον ίδιο τρόπο χρησιμοποιούν μια μετρική έτσι ώστε να ταξινομούν τις κλάσεις, μόνο που στη συγκεκριμένη περίπτωση η μετρική που χρησιμοποιούν δεν είναι γραμμική (12).

Τα Σχεσιακά Δίκτυα ακολουθούν την ίδια διαδικασία με τα Πρωτότυπα Δίκτυα όπου υπολογίζονται για την κάθε κλάση τα χαρακτηριστικά της και αθροίζονται μεταξύ τους για να φτιάξουν κάποιου είδους πρωτότυπο. Στη συνέχεια, η κάθε εικόνα περνάει από μια μη γραμμική συνάρτηση η οποία της δίνει ένα σκορ από το 0 εώς το 1 για το αν ανήκει σε μια συγκεκριμένη κλάση (12). Ουσιαστικά λοιπόν η διαφορά τους από τα δύο προηγούμενα νευρωνικά δίκτυα είναι ότι το δίκτυο μαθαίνει μια μη γραμμική μετρική και τη χρησιμοποιεί χωρίς να έχει μια δεδομένη μετρική από την αρχή της διαδικασίας. Η διαφορά των τριών νευρωνικών δικτύων είναι εμφανής στην εικόνα (5).



Σχήμα 5: Η διαφορά των τριών διαφορετικών δικτύων

3.3.5 Δίκτυα Συνδιακύμανσης

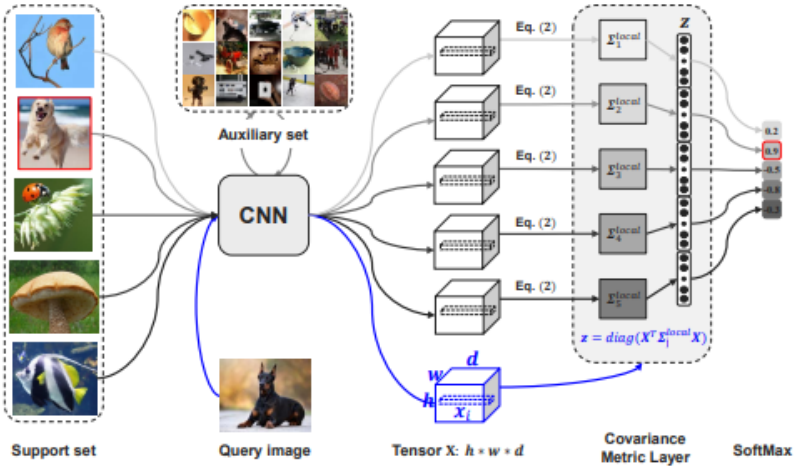
Τα Δίκτυα Συνδιακύμανσης (Covariance Network) είναι τα τελευταία νευρωνικά δίκτυα στα οποία θα γίνει αναφορά από την κατηγορία της μάθησης μέσω μετρικής. Τα συγκεκριμένα δίκτυα χρησιμοποιούν τα χαρακτηριστικά με την περισσότερη πληροφορία για να εξάγουν κανόνες για τις κλάσεις τους. Μέσα από αυτά τα χαρακτηριστικά

δημιουργούν μια αναπαράσταση για την κάθε κλάση. Με αυτό τον τρόπο ορίζουν μία μετρική για το διαχωρισμό των κλάσεων που βασίζεται στη συνδιακύμανση των εικόνων της κάθε κλάσης (13). Οπότε τα Δίκτυα Συνδιακύμανσης χρησιμοποιούν τα χαρακτηριστικά του προβλήματος για τη δημιουργία της μετρικής και μέσα από τον πίνακα συνδιακύμανσης που δημιουργούν παίρνουν πρώτης και δεύτερης τάξης πληροφορία για την κάθε κλάση (13).

Για περαιτέρω ανάλυση της διαδιακασίας, θα χρησιμοποιηθεί το παρακάτω παράδειγμα. Έστω ότι υπάρχει ένα σύνολο εικόνων k από την C κλάση ενός προβλήματος, τότε $D_C = (X_1, X_2, ..., X_k), X_i \in \mathcal{R}^{d*M}$, όπου d είναι η διάσταση του συνόλου D, όπου η συγκεκριμένη κλάση έχει k διαφορετικές εικόνες και η κάθε εικόνα έχει M χαρακτηριστικά, τότε η μετρική σε αυτό το πρόβλημα ορίζεται ως εξής:

$$\sum_C = \frac{\sum_{i=1}^{K}(X_i - \tau)(X_i - \tau)^T}{MK - 1}$$

Όπου το $\tau \in \mathcal{R}^{d*M}$ είναι ένας πίνακας διανυσμάτων μέσης τιμής για το κάθε χαρακτηριστικό των εικόνων της συγκεκριμένης κλάσης του προβλήματος. Ο παραπάνω τύπος δίνει το αποτέλεσμα που υπάρχει σε κάθε στοιχείο του πίνακα συνδιακύμανσης για κάθε χαρακτηριστικό και σε κάθε συγκεκριμένη κλάση C του προβλήματος. Οπότε για κάθε κλάση του προβλήματος δημιουργούνται τέτοιοι πίνκαες που αφορούν όλα τα χαρακτηριστικά της κάθε κλάσης ξεχωριστά. Για παράδειγμα, σε ένα πρόβλημα που έχουμε 6 διαφορετικές κλάσεις και για την κάθε κλάση υπάρχουν 6 παραδείγματα το ονομάζουμε 6 κλάσεις (way) 6 παραδείγματα (shot). Ακόμα αν γίνει η υπόθεση οτι τα χαρακτηριστικά για την κάθε εικόνα είναι M = 500. Αυτό σημαίνει ότι δημιουργούνται MK = 500 * 6 δείγματα συνολικά για μία κλάση. Οπότε στη συνέχεια χρησιμοποιούνται αυτά τα 3000 δείγματα για να υπολογιστεί ο πίνακας συνδιακύμανσης για την απεικόνιση της κλάσης. Άρα για κάθε κλάση υπάρχει ένας τέτοιος πίνακας που έχει δημιουργήσει 3000 δείγματα για να δωωθεί η πληροφορία που θα είναι ικανή να διαχωρίσει τις κλάσεις του προβλήματος.



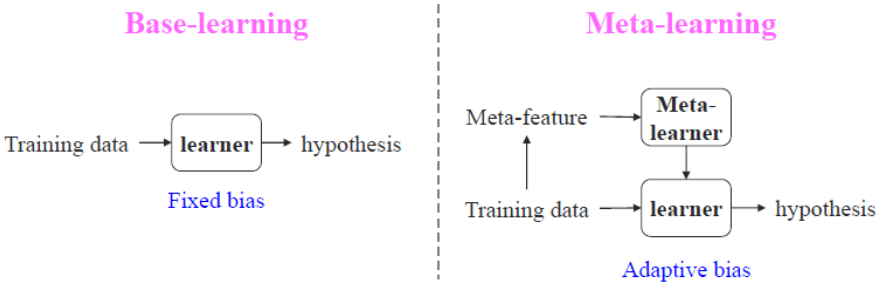
Σχήμα 6: Δίκτυα Συνδιακύμανσης

3.4 Μέθοδοι που βασίζονται στη μετα-μάθηση

Οι αλγόριθμοι μετα-μάθησης διαφέρουν αρκετά από τους αλγόριθμους μάθησης μέσω μετρικής, οι οποίοι βασίζονται στη μετρική που χρησιμοποιείται για το διαχωρισμό των κλάσεων του προβλήματος, ενώ οι αλγόριθμοι μετα-μάθησης βασίζονται στη δημιουργία μετα-δεδομένων.

Η λογική που τους διέπει είναι ότι αρχικά από τα δεδομένα εξάγουν μετα-δεδομένα, δηλαδή πληροφορίες που εκφράζονται μέσα από τα δεδομένα και δεν υπάρχουν εν γένει στα χαρακτηριστικά του κάθε προβλήματος. Στη συνέχεια αυτές οι πληροφορίες εισάγονται σε ένα μετα-εκπαιδευτή όπου εκπαιδεύεται πάνω σε αυτά τα δεδομένα και δημιουργούνται κάποια βάρη τα οποία εισάγονται στην κανονική εκπαιδευτική διαδικασία του μοντέλου, για να εξάγουν, μαζί με τα αρχικά δεδομένα, τα αποτελέσματα στο εκάστοτε πρόβλημα.

Οι μέθοδοι μετα-μάθησης φάνηκαν ιδιαίτερα χρήσιμες στη μάθηση με την χρήση λίγων δεδομένων. Η βασική ιδέα για να χρησιμοποιηθούν τέτοιες μέθοδοι ήταν να αξιοποιούνται συγχρόνως πολλά προβλήματα λίγων δεδομένων και να εξάγονται από αυτά πολλά μετα-δεδομένα τα οποία θα είναι κοινά για προβλήματα που έχουν κοινά χαρακτηριστικά έτσι ώστε να δίνεται βοήθεια στη βασική ροή του μοντέλου και να μπορεί να βγάζει συμπεράσματα με πολύ λίγα δεδομένα (14). Άρα, όπως αναφέρεται και στη εργασία (14), κάθε διαφορετικό πρόβλημα λίγων δεδομενων να μπορεί να χρησιμοποιεί τα μετα-δεδομένα και από άλλα προβλήματα λίγων δεδομένων με παρόμοια χαρακτηριστικά.



Σχήμα 7: μετα-Μάθηση

3.4.1 Μετα-μάθηση σε διαφορετικά πεδία

Η μετα-μάθηση σε διαφορετικά πεδία (Model Agnostic Meta Learning η αλλιώς MAML) είναι μια αρχιτεκτονική μετα-μάθησης η οποία πήρε το όνομα της από τη δυνατότητα εκπαίδευσης σε πολύ διαφορετικά προβλήματα τα οποία μπορεί να συνδυάσει και να εξάγει πληροφορία για το κάθε ένα ξεχωριστά. Η συγκεκριμένη μέθοδος μπορεί να είναι αποτελεσματική σε οποιοδήποτε μοντέλο χρησιμοποιεί μέθοδο βημάτων κλίσης (15). Οι αλγόριθμοι αυτής της κλάσης επιλέγουν ένα σύνολο από παραμέτρους, οι οποίες μπορούν να εκπαιδευτούν καλά για ένα άλλο πρόβλημα μετά από ένα, δύο βήματα μιας διαδικασίας βημάτων κλίσης. Με αυτή τη διαδικασία για ένα τέτοιο δίκτυο είναι εύκολο να μπορεί να μάθει από ένα σύνολο παραμέτρων το οποίο θα μπορεί να χρησιμοποιηθεί σε διαφορετικά προβλήματα. Συνεπώς οι αλγόριθμοι μετα-μάθησης εκπαιδευόμενοι σε διαφορετικά πεδία έχουν τη δυνατότητα να εκπαιδεύονται πολύ εύκολα ακόμα και σε προβλήματα όπου τα δεδομένα που υπάρχουν είναι όχι μόνο λίγα αλλά έχουν και πάρα πολλά κενά. Αυτή η ικανότητα τους βασίζεται στο ότι επιλέγουν σύνολο παραμέτρων που μπορούν να βοηθήσουν στην εκμάθηση άλλων πρόβλημάτων. Έτσι τους δίνεται μεγαλύτερη δυνατότητα γενίκευσης (15). Αυτή η δυνατότητα γενίκευσης βασίζεται στο ότι οι παράμετροι του μοντέλο μπορεί να ξεκινούν από ένα αρχικό σημείο και μετά από ένα βήμα κλίσης να διαφέρουν αρκετά από πρόβλημα σε πρόβλημα. Αυτό έχει σαν αποτέλεσμα να μπορούν αυτές οι παράμετροι να προσαρμόζονται ανάλογα το πρόβλημα που έχουν να αντιμετωπίσουν, οπότε η γενίκευση του μοντέλου επιτυχγάνεται μέσο της ικανότητας του, να αντιμετωπίσουν, οι παράμετροι του με μεγάλη ευκολία πολύ διαφορετικά προβλήματα.

Οι παράμετροι ενός τέτοιου δικτύου από τη στιγμή που το μοντέλο για να εκπαιδευτεί χρειάζεται ένα μικρό αριθμό από αλλαγές του βήματος κλίσης, οδηγούνται σε πολύ γρήγορη εκμάθηση για την κάθε κλάση ξεχωριστά. Ο τύπος που δίνεται για την κάθε παράμετρο είναι ο ακόλουθος (15):

$$\min_{\theta} \sum_{i} \mathcal{L}(\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, \mathcal{D}_i), \mathcal{T}_i)$$

όπου:

θ: η διανυσματική παράμετρος που εκτιμάται.

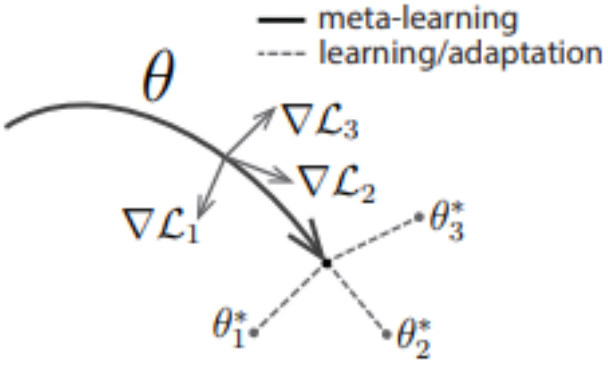
.

$\mathcal{D}_i$: Τα δεδομένα της εκπαίδευσης για την κλάση i.

$\mathcal{T}_i$: Τα δεδομένα του σύνολο που θα ελέγξουμε για την κλάση i.

Οπότε ο στόχος για την κάθε παράμετρο ϑ είναι σε κάθε βήμα να μειώνεται αυτή η διαφορά χρησιμοποιώντας σε κάθε βήμα την νέα ϑ που έχει δημιουργηθεί. Το μόνο μειονέκτημα σε αυτήν τη διαδικασία είναι ότι τόσο ο υπολογισμός της ϑ όσο και ο επανυπολογισμός της ϑ σε κάθε βήμα χρειάζεται την κλίση (gradient). Οπότε για να πάρουμε την αλλαγή της παραμέτρου σε κάθε βήμα χρειάζεται η δεύτερη παράγωγος της συνάρτησης του ϑ (δηλαδή πρέπει να χρησιμοποιηθεί ο Εσσιανός Πίνακας) (15).

Για να βελτιωθεί αυτό το πρόβλημα ο (15) πρότεινε τη πρώτης τάξης μετα-μάθηση σε διαφορετικά πεδία ή αλλιώς FOMAML (First Order Model Agnostic Meta Learning) στην οποία παραλείπεται ο υπολογισμός των δεύτερων παραγώγων. Όμως, αυτό το γεγονός δεν είχε ως αποτέλεσμα τη μείωση του ποσοστού επιτυχίας του αλγορίθμου. Αυτή η επιτυχία βασίστηκε όπως αναφέρει ο (15), στην χρήση Relu δικτύων τα οποία είναι σχεδόν γραμμικές συναρτήσεις οπότε οι δεύτερες παράγωγοι τους είναι σχεδόν μηδενικές. Τα Relu δίκτυα είναι φτιαγμένα για να συμπεριφέρονται με σχεδόν γραμμικό τρόπο και να είναι πιο εύκολο να βελτιστοποιούνται, οπότε περνούν οι παράμετροι του μοντέλου από μια Relu συνάρτηση αντί να υπολογίζουν τη δεύτερη παράγωγο σύμφωνα με (17).



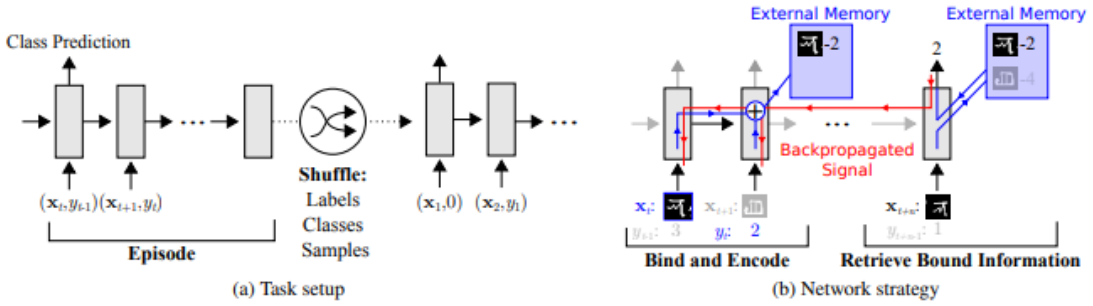
Σχήμα 8: μετα-μάθηση σε διαφορετικά πεδία

3.4.2 Νευρωνικά Δίκτυα αυξανόμενης μνήμης

Τα Νευρωνικά Δίκτυα Αυξανόμενης Μνήμης (Memory Augment Neural Network) πρώτο εμφανίστηκαν στο χώρο της μάθησης με λίγα δεδομένα με την εργασία του (18). Τα συγκεκριμένα δίκτυα εκπαιδεύονται για μία κλάση κάθε φορά. Δέχονται τα δεδομένα κάθε κλάσης σαν ένα ζεύγος από μια σειρά x και μία εξαρτημένη μεταβλητή y. Συνεπώς η εξαρτημένη μεταβλητή y τη χρονική στιγμή t δε μαθαίνει από τα δεδομένα x που έχουν παραχθεί μέχρι τη συγκεκριμένη χρονική στιγμή αλλά περιμένει μέχρι την επόμενη χρονική στιγμή t + 1 (18). Οπότε το σύστημα μαθαίνει το κάθε παράδειγμα από παλαιότερη πληροφορία. Τα δεδομένα ανακατεύονται κάθε φορά που αναπαρίσταται μία κλάση, με αυτό τον τρόπο το δίκτυο δε μαθαίνει τη σειρά με την οποία εμφανίζονται τα δεδομένα αλλά την σχέση που έχουν με την κλάση.

Το δίκτυο λοιπόν έχει έναν ελεγκτικό μηχανισμό ο οποίος αποθηκεύει πληροφορία και το βοηθά να θυμάται έτσι ώστε να κατηγοριοποιεί τα δεδομένα στις κλάσεις που ανήκουν. Στην πράξη αυτός ο ελεγκτικός μηχανισμός είναι είτε ένα δίκτυο μικρής μνήμης είτε ένα δίκτυο με προς τα εμπρός τροφοδότηση (18). Οι πληροφορίες που έχουν αποθηκευτεί για την κάθε κλάση, αποκτήθηκαν με τη βοήθεια μίας μετρικής. Η μνήμη που έχει το δίκτυο για την κάθε κλάση είναι το άθροισμα όλων των πληροφοριών που έχουν αποθηκευτεί για την κάθε κλάση μέσα από τις σχέσεις που έχουν δημιουργηθεί από τη μετρική.

Με αυτήν τη διαδικασία όταν μια νέα κλάση εισέρχεται στο δίκτυο, πρώτα το δίκτυο αποκτά πληροφορία για αυτή την κλάση, έτσι ώστε στη συνέχεια να μπορεί να προβλέψει και νέες παρατηρήσεις που θα εμφανιστούν από την ίδια κλάση. Περνώντας τη διαδικασία το δίκτυο θα αποκτά μεγαλύτερη ικανότητα στο να μπορεί να προβλέψει τη συγκεκριμένη κλάση. Ακόμα, κατά τη διάρκεια της μετα-εκπαίδευσης μαθαίνονται οι παράμετροι του δικτύου, ώστε να δουλεύει καλύτερα για περισσότερες κλάσεις (18).



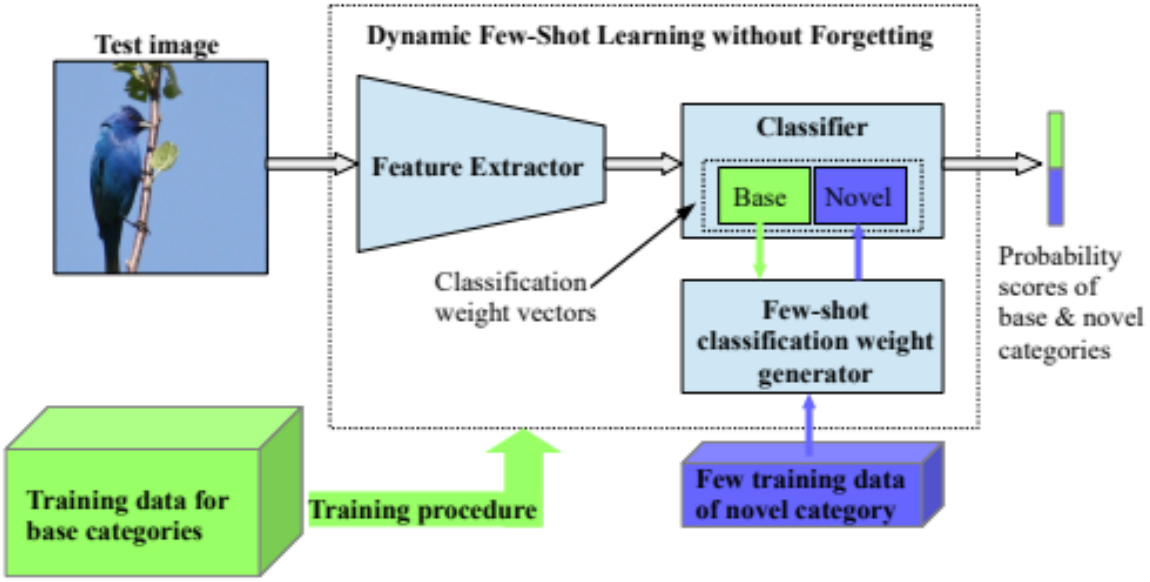
Σχήμα 9: Νευρωνικά Δίκτυα αυξανόμενης μνήμης

3.4.3 Δυναμική μάθηση με λίγα δεδομένα

Τα Δυναμικά Νευρωνικά Δίκτυα (Dynamic Neural Network) είναι μια ακόμα κατηγορία νευρωνικών δικτύων που ανήκει στην οικογένεια της μετα-μάθησης. Η δομή του δικτύου αποτυπώνεται στην ονοματοδοσία του αφού αποτελεί μια δυναμική διαδικασία. Η αρχιτεκτονική του διαφέρει από τα προηγούμενα δίκτυα που έχουν αναφερθεί γιατί συνεχίζεται η εκμάθηση κατά τη διάρκεια της παραγωγής αποτελεσμάτων από το σύνολο δεδομένων ελέγχου του μοντέλου (19).

Αρχικά, στο δίκτυο υπάρχει ένας ταξινομητής που αναγνωρίζει τα βασικά χαρακτηριστικά των δεδομένων και τα μετα-χαρακτηριστικά τους. Σε αυτό τον ταξινομητή τοποθετούνται αντίστοιχα βάρη για τα βασικά χαρακτηριστικά των δεδομένων και δημιουργείται παράλληλα και μία μηχανή δημιουργίας βαρών για τα μετα-χαρακτηριστικά (19). Η συγκεκριμένη μηχανή παίρνει πληροφορία από τα βασικά χαρακτηριστικά των δεδομένων και από τα μετα-χαρακτηριστικά. Η εκμάθηση των μετα-χαρακτηριστικών συνεχίζει κατά τη διάρκεια των αποτελεσμάτων του μοντέλου.

Η Δυναμική μάθηση με λίγα δεδομένα έχει χρησιμοποιήθει σε διάφορες εφαρμογές. Ο σκοπός της όπως αναφέρεται στο (19) είναι κατά τη διάρκεια των αποτελεσμάτων να μαθαίνει πληροφορία για τα μετα-χαρακτηριστικά χωρίς να αλλοιώνετε η εκμάθηση που έχει γίνει νωρίτερα. Στο τέλος της διαδικασίας το δίκτυο βγάζει για το κάθε χαρακτηριστικό μία πιθανότητα για το σε ποια κατηγορία ανήκει, ταξινομώντας με αυτό τον τρόπο τις εικόνες στην κατηγορία που ανήκουν σύμφωνα με το σταθμισμένο αποτέλεσμα του μοντέλου (19).



Σχήμα 10: Δυναμική μάθηση με λίγα δεδομένα

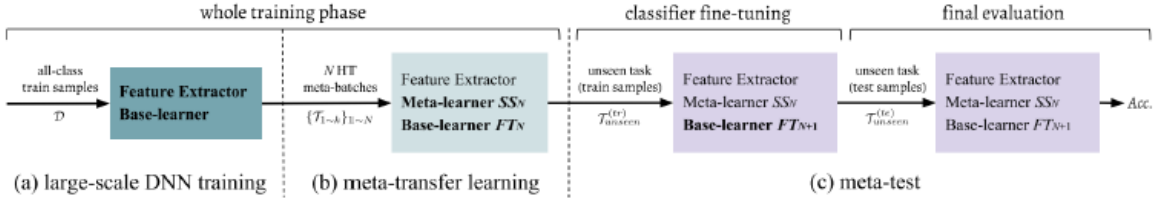
3.4.4 Μετα-μάθηση με μεταφορά

Η μετα-μάθηση με μεταφορά (Meta Transfer Learning) είναι μια ιδιαίτερα διαδεδομένη τεχνική στο χώρο της μάθησης με λίγα δεδομένα αφού χρησιμοποιείται σε πολλές περιπτώσεις (20). Η διαδικασία της εκμάθησης διαφέρει αρκετά από τους προηγούμενους αλγορίθμους που αναφέρθηκαν στο χώρο της μετα-μάθησης.

Η αρχιτεκτονική της μετα-μάθησης με μεταφορά αποτελείται από τρία στάδια για την εκμάθηση των χαρακτηριστικών του σύνολο εκπαίδευσης (20). Στο πρώτο στάδιο εκπαιδεύεται σε ένα μεγάλο σύνολο δεδομένων έτσι ώστε να παραχθεί ο τρόπος με τον οποίο θα επιλέγονται τα χαρακτηριστικά του προβλήματος από το βασικό μηχανισμό μάθησης. Η διαδικασία ξεκινάει σε ένα πολύ μεγάλο σύνολο δεδομένων που θα αφορά και δεδομένα που δεν υπάρχουν στο αρχικό σύνολο που θα εκπαιδευτεί στη συνέχεια. Αυτό γίνεται έτσι ώστε να αποκτήσει περισσότερη πληροφορία για το πρόβλημα και να μπορεί να την χρησιμοποήσει για να καταλαβαίνει καλύτερα την αξία των χαρακτηριστικών του προβλήματος (20).

Το επόμενο βήμα της εκπαίδευσης είναι η εισαγωγή των δεδομένων από το σύνολο εκπαίδευσης από όλα τα προβλήματα της διαδικασίας ώστε να προσαρμοστεί ο βασικός μηχανισμός μάθησης σε όλα τα προβλήματα της διαδικασίας. Στη συνέχεια εισάγονται τα δεδομένα του κάθε προβλήματος ξεχωριστά όπου εδώ το μοντέλο εκπαιδεύεται έχοντας την πρωταρχική πληροφορία από τα μετα δεδομένα που δημιουργήθηκαν και από τα δεδομένα μεγαλύτερης κλίμακας που είχαν μεταφερθεί αρχικά από το βασικό

μηχανισμό μάθησης (20). Το τελικό βήμα είναι όπως και σε όλους τους άλλους αλγορίθμους η αξιολόγηση των αποτελεσμάτων και η ταξινόμηση τους σε κλάσεις ανάλογα με το σταθμισμένο ποσοστό που έχουν πάρει από την τελική αξιολόγηση.



Σχήμα 11: μετα-μάθηση με μεταφορά

4 Μπεϋζιανή Μετα-μάθηση

4.1 Εισαγωγή

Τα τελευταία χρόνια η επιτυχία της Βαθιάς Μηχανικής Μάθησης σε πολλούς τομείς όπως η αναγνώριση εικόνων ή η μετάφραση μηχανών είναι αναμφισβήτητη. Στις περισσότερες από αυτές τις εφαρμογές υπάρχει μεγάλος όγκος δεδομένων που βοηθάει αυτές τις τεχνικές να αναπτύσσονται και να δίνουν καλύτερα αποτελέσματα. Αυτές οι τεχνικές όμως δοκιμαζόμενες σε προβλήματα με λιγότερα δεδομένα απέτυχαν να γενικεύσουν. Για αυτό το λόγο δημιουργήθηκε η τεχνική της μετα-μάθησης (29) και (30). Η συγκεκριμένη τεχνική για να είναι πιο αποτελεσματική χρησιμοποιεί μια στρατηγική που ονομάζεται επεισοδιακή εκπαίδευση (episodic training strategy) στην οποία εκπαιδευείς και αξιολογείς τα αποτελέσματα για το κάθε πρόβλημα ξεχωριστά. Με τον όρο επεισοδιακή αναφέρεται στο κάθε πρόβλημα σαν επεισόδιο.

Η μετα-μάθηση έχει ως σκοπό την αντιμετώπιση της έλλειψης δεδομένων στους τομείς που χρησιμοποιείται. Η αντιμετώπιση αυτή επιτυχγάνεται με την κατασκευή μοντέλων τα οποία εκπαιδεύονται σε μια κατανομή προβλημάτων και όχι μόνο σε ένα συγκεκριμένο πρόβλημα. Η διαδικασία αυτή έχει ως αποτέλεσμα το μοντέλο να εκπαιδεύεται σε όλα τα δεδομένα από όλα τα προβλήματα και να παίρνει πληροφορία την οποία θα μπορεί να χρησιμοποιήσει στην συνέχεια στο κάθε πρόβλημα ξεχωριστά. Ο μηχανισμός αυτός λοιπόν είναι υπεύθυνος για το όνομα της τεχνικής της μετα-μάθησης αφού μετακυλίεται η πληροφορία από άλλα σετ δεδομένων στο πρόβλημα για την εξαγωγή συμπερασμάτων.

Η διαδικασία της μετα-μάθησης ξεκινάει με το επίπεδο της εκπαίδευσης (meta-training). Στην εκπαίδευση του μοντέλου χρησιμοποιούνται παραπάνω από ένα σετ δεδομένων τα οποία εκπαιδεύονται στις ίδιες παραμέτρους και δίνουν πληροφορία στο μοντέλο, κάποιες φορές αρκετά ετερόκλητη όπως στην περίπτωση της μετα-μάθησης εκπαιδευμένη σε διαφορετικά πεδία (MAML). Τα δεδομένα εκπαιδεύονται με κοινές παραμέτρους οι οποίες μετά από κάποια βήματα κλίσης αλλάζουν και δίνουν πληροφορία συνολική στο επίπεδο της μετα-εκπαίδευσης η οποία μπορεί να χρησιμοποιηθεί ανάλογα στο κάθε πρόβλημα. Συγχρόνως όμως εκπαιδεύονται και στα δικά τους δεδομένα μόνο και αποκτούν πληροφορία που όπως θα δούμε στην συνέχεια χρησιμοποιούν ανάλογα με τη δομή του κάθε προβλήματος.

Έχοντας χρησιμοποιήσει αρκετά σετ δεδομένων στο επίπεδο εκπαίδευσης το μοντέλο πηγαίνει στο πρώτο επίπεδο αξιολόγησης (meta-validation) στο οποίο χρησιμοποιούνται δεδομένα από τα σετ δεδομένων που είχαν χρησιμοποιηθεί στο επίπεδο εκπαίδευσης και δεν τα είχε δει το μοντέλο. Στο τελευταίο επίπεδο ελέγχου (meta-test) το μοντέλο έρχεται αντιμέτωπο με σετ δεδομένων τα οποία είτε προέρχονται από σετ δεδομένων που εμφανίζονται στο σετ εκπαίδευσης είτε είναι τελείως άγνωστα στο

μοντέλο και τα αντιμετωπίζει πρώτη φορά. Σε όλη αυτή τη διαδικασία το μοντέλο πρέπει να αντιμετωπίσει ένα πρόβλημα που αποτελείται από πολλά διαφορετικά σετ δεδομένων κάθε ένα από τα οποία είναι ένα υποπρόβλημα (task) ή πρόβλημα μικρότερης κλίμακας.

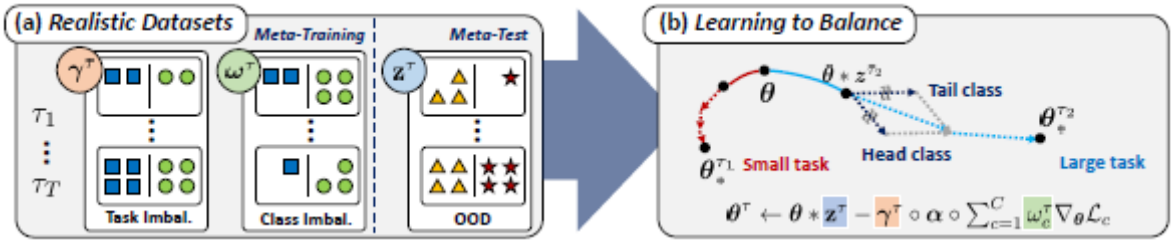
Στην συγκεκριμένη διπλωματική εργασία θα μελετήσουμε τη μετα-μάθηση κυρίως υπό το πρίσμα της Μπεϋζιανής Στατιστικής. Η Μπεϋζιανή Στατιστική μπορεί να χρησιμοποιηθεί σαν εργαλείο στο χώρο της Μηχανικής Μάθησης με τον ίδιο τρόπο που έχει χρησιμοποιηθεί και η Κλασσική Στατιστική. Η Μπεϋζιανή μετα-μάθηση (Bayesian Meta Learning) είναι ο συνδυασμός της Μπεϋζιανής Στατιστικής με τους αλγόριθμους της μάθησης με λίγα δεδομένα. Υπάρχουν διάφορες προσεγγίσεις που χρησιμοποιείται η Μπεϋζιανή Μάθηση στο χώρο της Μηχανικής Μάθησης (21), (22), (24).

Στη Μπεϋζιανή Στατιστική μπορούν να υπάρξουν διαφορές από τη δομή του κάθε δικτύου μέχρι τον τρόπο ταξινόμησης των αποτελεσμάτων. Η βασική διαφορά της με τις άλλες τεχνικές είναι ότι οι παράμετροι του εκάστοτε μοντέλου θεωρούνται ως τυχαίες μεταβλητές που ακολουθούν κάποια κατανομή. Αυτή η διαφορά δίνει τη δυνατότητα στη μετα-μάθηση, χρησιμοποιώντας το πλαίσιο της Μπεϋζιανής Στατιστικής, να αντιμετωπίσει σαν τυχαίες μεταβλητές τις παραμέτρους οι οποίες θα αλλάζουν πολύ εύκολα για να μπορούν να προσαρμοστούν σε ένα σύνολο προβλημάτων τα οποία είναι πολύ ετερόκλητα μεταξύ τους(που όπως αναφέραμε παραπάνω είναι από τις βασικές επιδιώξεις της μετα-μάθησης).

4.2 Προκλήσεις στη μάθηση με λίγα δεδομένα

Στις περισσότερες εφαρμογές της μετα-μάθησης με λίγα δεδομένα (25) το σενάριο του κάθε προβλήματος περιλαμβάνει ένα πρόβλημα ταξινόμησης με πολλές διαφορετικές κλάσεις στις οποίες ο αριθμός των δεδομένων στην εκπαίδευση του μοντέλου για την κάθε κλάση είναι ο ίδιος. Αυτός είναι ένας πάρα πολύ αυστηρός περιορισμός σε ένα πραγματικό σενάριο. Στην πραγματικότητα τα παραδείγματα που υπάρχουν για διαφορετικά υπο-προβλήματα ενός προβλήματος μπορεί να διαφέρουν (ανισσοροπία προβλημάτων) και τα παραδείγματα για τις κλάσεις ενός προβλήματος μπορεί να διαφέρουν πολύ από κλάση σε κλάση (ανισσοροπία κλάσεων). Για παράδειγμα για ένα πρόβλημα μπορεί να έρθει ένα νέο υποπρόβλημα που έχει προκύψει από διαφορετική κατανομή από αυτή που έχει εκπαιδευτεί το μοντέλο. Παραθέτουμε παρακάτω μία είκονα που περιγράφει αυτή ακριβώς την περίπτωση (25).

Ακόμα σε ένα ρεαλιστικό σενάριο, η μετα-μάθηση που μπορούμε να αντλήσουμε από διαφορετικά προβλήματα απέχει πάρα πολύ (25).Είναι μια λογική συνέπεια και αφορά τον ορισμό και τη δομή του κάθε προβλήματος. Ένα πρόβλημα με πολύ μεγάλο αριθμό κλάσεων και λίγα δεδομένα για κάθε κλάση μπορεί να συνεισφέρει έναν αρκετά μεγάλο όγκο πληροφορίας συνολικά αλλά λιγότερο για την κάθε κλάση ξεχωριστά. Από την άλλη ένα πρόβλημα με λίγες κλάσεις και πολλά δεδομένα για την κάθε κλάση μπορεί



Σχήμα 12: Μη ισορροπημένο σύνολο δεδομένων

να δώσει καλύτερες κατευθύνσεις για τις κλάσεις που έχει. Σε κάποια προβλήματα συνηθίζεται να διαφέρει ο αριθμός των παρατηρήσεων ανάλογα με την κλάση του προβλήματος.

Το γεγονός αυτό έχει ως αποτέλεσμα κατά τη διαδικασία εκπαίδευσης να διαφέρει το πόσο χρήσιμη είναι η πληροφορία της μετα-μάθησης που συγκεντρώνεται για όλα τα προβλήματα μαζί σε σχέση με την πληροφορία που έχει το κάθε πρόβλημα ξεχωριστά από το δικό του σετ δεδομένων εκπαίδευσης. Προβλήματα με μικρό αριθμό δεδομένων εκπαίδευσης ή προβλήματα που εκπαιδεύονται στο επίπεδο μετα-εκπαίδευσης, βασίζονται κυρίως στη μετα-μάθηση που αποκτούν σε σχέση με άλλα προβλήματα που έχουν μια κατανομή στα δεδομένα τους ή έχουν μεγαλύτερο αριθμό δεδομένων εκπαίδευσης. Τέτοια προβλήματα με μεγάλο αριθμό δεδομένων εκπαίδευσης θα έχουν ένα καλύτερο αποτέλεσμα όταν εκπαιδεύονται κυρίως στον κύριο μηχανισμό εκπαίδευσης.

Στα προβλήματα ταξινόμησης πολλών κλάσεων είναι σημαντικό να γίνει προσπάθεια διαχείρισης της διαδικασίας μάθησης της κάθε κλάσης διαφορετικά για να αντιμετωπιστεί το πρόβλημα των μη ισορροπημένων δεδομένων. Συνεπώς, για να αντιμετωπιστεί με βέλτιστο τρόπο η διαδικασία της μετα-μάθησης, λαμβάνοντας υπόψη και τα μη ισορροπημένα δεδομένα, είναι ωφέλιμο να μπορεί να αποφασίζει το μοντέλο πόσο θα προσαρμόζει την τιμή της κάθε παραμέτρου για την κάθε κλάση, του κάθε υποπρόβλήματος, από το σύνολο προβλημάτων της εκπαίδευσης.

Όλες αυτές οι αποφάσεις που χρειάζεται να παρθούν σε ένα τέτοιο πρόβλημα οδηγούν στο συμπέρασμα ότι μπορεί να υπάρχει πολύ μεγάλη αβεβαιότητα πάνω σε αυτές τις αποφάσεις και στο κατά πόσο είναι σωστές να παρθούν. Οπότε, το παραπάνω πρόβλημα έχει ένα μεγάλο ποσοστό αβεβαιότητας . Η δομή ενός Μπεϋζιανού δικτύου βασίζεται στην ύπαρξη αβεβαιότητας για το μοντέλο που θα κατασκευαστεί, γεγονός που είναι μια από τις βασικές αρχές στη χρήση της μάθησης με λίγα δεδομένα, ειδικότερα και από τη στιγμή που αρχίζει να χρησιμοποιείται όλο και περισσότερο σε εφαρμογές για αναγνώριση προσώπου και διαγνώσεις φαρμάκων (25).

Οι περισσότερες μπεϋζιανές προσεγγίσεις της μάθησης με λίγα δεδομένα αντιμετωπίζουν τις παραμέτρους του προβλήματος σαν τυχαίες μεταβλητές. Αυτές οι τυχαίες μεταβλητές

ακολουθούν συγκεκριμένες κατανομές. Οι κατανομές αυτές σε πολύ σπάνιες περιπτώσεις μπορεί να είναι γνωστές κατανομές όπως η κανονική κατανομή ή η ομοιόμορφη αλλά κάτι τέτοιο είναι δύσκολο να συμβεί σε πραγματικά προβλήματα. Αυτό, έχει σαν αποτέλεσμα να χρησιμοποιούνται κατανομές που είτε είναι αποτέλεσμα προσεγγιστικών κατανομών (33), (38) ή διαμορφώνονται μέσω της χρήση πολλαπλών διαφορετικών βαρών για το ίδιο δείγμα (39), (50). Αυτές οι κατανομές προσπαθούν να προσεγγίσουν την συμπεριφορά της παραμέτρου και να την κατατάξουν σε κάποια από τις γνωστές κατανομές. Η πρακτική των προσεγγιστικών κατανομών αντιμετωπίζει προβλήματα στις εκτων υστέρων κατανομές. Σε μια τέτοια προσέγγιση οι εκ των υστέρων κατανομές περιορίζονται πολύ στην εκφραστικότητα τους με αποτέλεσμα να δίνουν κατανομές με λίγη πληροφορία. Το αποτέλεσμα σε αυτή την περίπτωση είναι να μην γίνονται μεγάλες αλλαγές στις παραμέτρους του προβλήματος ανάλογα με το σετ δεδομένων που θα χρησιμοποιηθεί.

Η τεχνική της χρήση πολλαπλών διαφορετικών βαρών είναι πολύ αργή υπολογιστικά και κοστοβάρα όσο αναφορά τον αποθηκευτικό χώρο. Όμως, στην περίπτωση χρήσης εκ των προτέρων κατανομών που να έχουν σημασία στο χώρο ενός προβλήματος, είναι γνωστό ότι θα αντιμετωπιστούν δυσκολίες λόγω των περίπλοκων σχέσων μεταξύ των βαρών και των συναρτήσεων που υπάρχουν σε ένα βαθύ νευρωνικό δίκτυο (39). Με αυτή την πρακτική θα γίνει πιο εύκολη η αντιμετώπιση του προβλήματος που θα υπήρχε αν επιλεγόταν μια προσεγγιστική κατανομή που δεν θα επέτρεπε την αντίστοιχη ελευθερία στην επιλογή της κατανομής ανάλογα με το υπόπροβλημα που θα έρθει αντιμέτωπη.

Η βασική τεχνική που θα χρησιμοποιηθεί στη συγκεκριμένη διπλωματική εργασία θα είναι ένα δίκτυο Μπεϋζιανής μετα-μάθησης, το οποίο θα αναφέρεται ως Μπεϋζιανή προσαρμοσμένη μετα-μάθηση (Bayesian TAML ή (Task-Adaptive Meta-Learning) το οποίο μαθαίνει τις κατανομές των παραμέτρων για τις μεταβλητές του προβλήματος και προσπαθεί να ισορροπήσει το αντίκτυπο της μετα-μάθησης και της βασικής μάθησης του κάθε προβλήματος στα δεδομένα του.

Το όνομα αυτού του δικτύου οφείλεται στην αλλαγή των παραμέτρων ανάλογα με το σετ δεδομένων του προβλήματος. Στην αρχή λοιπόν της μετα-εκπαίδευσης παίρνουμε μια πρώτη αναπαράσταση για το κάθε πρόβλημα. Αυτή η πρώτη αναπαράσταση γίνεται μέσω της εκμάθησης της κατανομής του σύνολο δεδομένων και μερικών βασικών χαρακτηριστικών όπως της μέσης τιμής, της διασποράς και τον αριθμό των δεδομένων που υπάρχουν για την κάθε κλάση. Μετά από αυτή την εκμάθηση, η διαδικασία επικεντρώνεται στην εκμάθηση τριών σημαντικών παραμέτρων του δικτύου (25):

1) πολλαπλασιαστής ρυθμού μάθησης: Ο συγκεκριμένος πολλαπλασιαστής θα εξαρτάται από την κάθε κλάση. Η κάθε κλάση θα έχει διαφορετικό πολλαπλασιαστή έτσι ώστε σε κάποιες περιπτώσεις να είναι πολύ πιο κοντά στην πληροφορία που υπάρχει

από τη μετα-μάθηση και σε άλλες περιπτώσεις να μπορεί να παρεκκλίνει γρηγορότερα από την αρχική πληροφορία.

2) ποσοστό μάθησης: που εξαρτάται από την κλάση, το οποίο αποφασίζει πόσες πληροφορίες θα χρησιμοποιηθούν από κάθε κλάση, γιατί μπορεί τα μεγέθη των κλάσεων να διαφέρουν πολύ μεταξύ τους ακόμα και στο ποσοστό που βρίσκονται μέσα στα προβλήματα.

3) Ένας διαμορφωτής που εξαρτάται από τις αρχικές παραμέτρους και κλάσεις του κάθε προβλήματος. Σε ένα συγκεκριμένο πρόβλημα μπορεί να μην υπάρχει καθόλου μία κλάση του προβλήματος για αυτό θα πρέπει να υπάρχει κάποια τροποποίηση στα συγκεκριμένα προβλήματα σε σχέση με την αρχική κατανομή. Αν για όλα τα προβλήματα υπάρχει η ίδια κατανομή τότε θα υπάρχει μεγάλη ασυμφωνία με τα επιμέρους δεδομένα που θα έχει το κάθε πρόβλημα.

Αυτές οι τρεις παράμετροι στις οποίες θα αναφερθούμε και στην συνέχεια θα είναι η κύρια διαφορά της Μπεϋζιανής προσαρμοσμένης μετα-μάθησης που θα αναπτυχθούν σε σχέση με άλλες τεχνικές που θα γίνει αντιπαραβολή. Οι συγκεκριμένες παράμετροι θα αντιμετωπιστούν σαν τυχαίες μεταβλητές και θα υπολογίζονται για το κάθε πρόβλημα ξεχωριστά. Αυτές οι παράμετροι θα είναι ένας από τους λόγους που η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση είναι κατάλληλη για προβλήματα που εμφανίζεται ανισορροπία στα δεδομένα.

4.3 Εφαρμογές της μετα-μάθησης

Η μετα-μάθηση είναι ένας σύγχρονος τρόπος μάθησης μοντέλων,ωπου είναι χρήσιμος για προβλήματα με λίγα δεδομένα. Το μοντέλο που κατασκευάζεται, εκπαιδεύεται σε μια κατανομή από προβλήματα και αυτή η εκπαίδευση επιτυγχάνεται με μεθόδους που είναι βασισμένες είτε στη μνήμη, είτε στη μετρική, είτε στη βελτιστοποίηση των παραμέτρων του μοντέλου(29). Μέθοδοι οι οποίες βασίζονται στη μνήμη του μοντέλου αποθηκεύουν τα δεδομένα κάθε κλάσης μαζί και στην συνέχεια τα χρησιμοποιούν σε κάθε πρόβλημα ξεχωριστά όπως αναφέρει (18).

Μέθοδοι οι οποίες βασίζονται στη μετρική, αναφέρθηκαν και στο προηγούμενο κεφάλαιο μερικές τέτοιες περιπτώσεις και τι κάνει η κάθε μέθοδος ξεχωριστά, έχουν ως στόχο να ορίσουν την απόσταση μεταξύ των παρατηρήσεων. Στην συνέχεια , αυτές οι μέθοδοι ορίζουν την απόσταση της κάθε παρατήρησης από το πρωτότυπο της κάθε κλάσης που έχουν δημιουργήσει. Η απόσταση της κάθε παρατήρησης από το κάθε πρωτότυπο της κάθε κλάσης είναι ο ταξινομητής του προβλήματος. Τα δεδομένα ταξινομούνται στην κλάση από την οποία η κάθε παρατήρηση έχει τη μικρότερη απόσταση από το πρωτότυπο της συγκεκριμένης κλάσης (9), (10).

Η τρίτη κατηγορία από μεθόδους που χρησιμοποιούνται στη μετα-μάθηση βασίζεται

στη βελτιστοποίηση παραμέτρων του προβλήματος. Το δίκτυο έχει παραμέτρους οι οποίες είναι ικανές μετά από έναν αριθμό βημάτων κλίσης να είναι προσεγγιστικά οι βέλτιστες για όλη την κατανομή προβλημάτων που αφορά το συνολικότερο πρόβλημα (15). Η τρίτη κατηγορία είναι και η βασική που θα χρησιμοποιηθεί στην συγκεκριμένη εργασία και θα συγκριθεί με τις υπόλοιπες μεθόδους. Στη μετα-μάθηση σε διαφορετικά πεδία προτείνεται μια βελτιωμένη έκδοση της (13). Στην συγκεκριμένη εργασία έγινε προσπάθεια να παρατηρηθεί αν το μοντέλο θα έχει καλύτερα αποτελέσματα, μεταβάλοντας το ποσοστό εκμάθησης (learning rate) ανάλογα με το πρόβλημα της διαδικασίας. Η μετα-μαθηση σε διαφορετικά πεδία με μεταβαλλόμενο ποσοστό εκμάθησης θα είναι ένα από τα τρία μοντέλα που θα χρησιμοποιηθεί.

4.4 Τα προβλήματα της μετα-Μάθησης σε διαφορετικά πεδία

Η μετα-Μάθηση σε διαφορετικά πεδία είναι χρήσιμη σε πολλά διαφορετικά προβλήματα, αλλά αντιμετωπίζει μερικούς περιορισμούς από την αρχιτεκτονική της. Στη συγκεκριμένη εργασία θα γίνει χρήση της προσέγγισης του (15). Αρχικά υποθέτουμε την κατανομή ενός προβλήματος $p(\tau)$ από το σύνολο των προβλημάτων της διαδικασίας, η οποία γεννά τυχαία δεδομένα εκπαίδευσης X^τ, Y^τ και τυχαία δεδομένα ελέγχου $\tilde{X^\tau}, \tilde{Y^\tau}$. Στην συνέχεια, ο στόχος της μετα-μάθησης σε διαφορετικά πεδία είναι να δημιουργήσει μετα-γνώση για όλα τα προβλήματα της διαδικασίας και να δημιουργήσει πολλές παραμέτρους. Θα γίναι αναφορά σε μία παράμετρο ϑ η οποία να μπορεί να γενικεύει ανεξάρτητα από την κατανομή του κάθε προβλήματος έτσι ώστε να είναι εύκολο να δίνευαι το επόμενο βήμα αυτής της παραμέτρου για το κάθε πρόβλημα ξεχωριστά που θα οδηγήσει στη βέλτιστη λύση μετά από ένα ή και περισσότερα βήματα. Ο τύπος στον οποίο βασίζεται ο (15) είναι ο παρακάτω:

$$\min_{\theta} \sum_{i} \mathcal{L}(\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, \mathcal{D}_\rangle), \tilde{\mathcal{D}}_\rangle)$$

όπου στο συγκεκριμένο τύπο το α είναι το βήμα που μειώνεται το εμπειρικό λάθος $\mathcal{L}$. Ακόμα μέσω της μετα-μάθησης το ϑ μπορεί να μειώσει το εμπειρικό λάθος $\mathcal{L}$ έστω και αν ο αριθμός των δεδομένων είναι πάρα πολύ μικρός.

Ο παραπάνω τύπος εμφανίζει το πρώτο βήμα της διαδικασίας αλλά αυτή η διαδικασία ακολουθείται επαναληπτικά μέχρι να βρεθεί το ελάχιστο για την παράμετρο ϑ. Η συγκεκριμένη διαδικασία γίνεται για όλες τις παραμέτρους ενός τέτοιου μοντέλου όπου τα βήματα κλίσης που θα χρειαστούν για την κάθε παράμετρο μπορεί να διαφέρουν μεταξύ τους. Παρόλο της χρησιμότητας τους τα δίκτυα μετα-μάθησης σε διαφορετικά πεδία έχουν τους παρακάτω περιορισμούς, που δεν τους επιτρέπουν να αντιμετωπίζουν σύνολο προβλημάτων που έχουν διαφορετικό αριθμό δεδομένων στις κλάσεις ενός

προβλήματος, είτε σε όλα τα προβλήματα μιας διαδικασίας μετα-μάθησης, είτε όταν μερικές κλάσεις δεν υπάρχουν καθόλου σε μερικά προβλήματα της διαδικασίας (25):

1) Ανισσοροπία κλάσεων. Τα Δίκτυα μετα-μάθησης σε διαφορετικά πεδία δεν μπορούν να αντιμετωπίσουν την ανισορροπία των κλάσεων σε κάθε πρόβλημα ξεχωριστά. Σε περιπτώσεις, που σε ένα πρόβλημα μια κλάση έχει μεγάλο αριθμό δεδομένων εκπαίδευσης φαίνεται ότι κατά τη διάρκεια των βημάτων κλίσης του κάθε προβλήματος ξεχωριστά, εκπαιδεύεται κυρίως σε αυτή την κλάση οπότε έχει πολύ χαμηλά ποσοστά επιτυχίας στις άλλες κλάσεις. Αυτή η ανισορροπία δεν θα μπορούσε να αντιμετωπιστεί αφού όπως είδαμε παραπάνω στην κατασκευή των βέλτιστων παραμέτρων ϑ δεν υπάρχει καμία διαφοροποίηση ανάλογα με το πόσες παρατηρήσεις έχει κάθε κλάση ενός συγκεκριμένου προβλήματος.

2) Ανισσοροπία δεδομένων μεταξύ των προβλημάτων. Τα συγκεκριμένα Δίκτυα έχουν ένα δεδομένο αριθμό βημάτων κλίσης ο οποίος είναι ο ίδιος για όλα τα προβλήματα της διαδικασίας της μετα-μάθησης. Το γεγονός αυτό περιορίζει τη διαδικασία ώστε να μη μπορεί το δίκτυο να μάθει περισσότερο από κάποιο πρόβλημα της διαδικασίας αναλογικά με τα δεδομένα εκπαίδευσης που υπάρχουν για το συγκεκριμένο πρόβλημα. Αυτό το πρόβλημα θα μπορούσε να αντιμετωπιστεί με την προσθήκη κάποιων βαρών στην παραπάνω διαδικασία κάτι που δεν είναι εμφωλευμένο στη διαδικασία των μοντέλων της μετα-μάθησης σε διαφορετικά πεδία.

3) Κλάσεις που δεν υπάρχουν σε κάποια προβλήματα. Ενώ θεωρητικά μπορούν να αντιμετωπίσουν τις συγκεκριμένες περιπτώσεις, φαίνεται ότι η αντιμετώπιση που έχουν σε κλάσεις που δεν υπάρχουν σε κάποια προβλήματα της διαδικασίας είναι τελείως διαφορετική από τις κλάσεις που υπάρχουν σε όλα τα προβλήματα της διαδικασίας. Ακόμα μια παράμετρος, η οποία παίρνει πληροφορία από προβλήματα που δεν έχουν την συγκεκριμένη κλάση σίγουρα θα είναι λιγότερο σημαντική για την εκπαίδευση αυτής της κλάσης όταν μάλιστα την αντιμετωπίζει με τον ίδιο τρόπο όπως κλάσεις που προϋπήρχαν στην κατανομή των προβλημάτων.

Αυτοί οι τρεις λόγοι είναι και οι βασικοί λόγοι που θα χρησιμοποηθεί στην συγκεκριμένη διπλωματική εργασία σαν βασική τεχνική η προσαρμοσμένη μετα-μάθηση. Θα αναπτυχθούν και μοντέλα που θα αφορούν τη μετα-μάθηση σε διαφορετικά πεδία για να συγκριθούν τα αποτελέσματα τους σε πιο ρεαλιστικά προβλήματα. Η μετα-μάθηση σε διαφορετικά πεδία είναι μια πάρα πολυ καλή τεχνική με πολλά πλεονεκτήματα και πολλές εφαρμογές όπως αναφέθηκε και στο Κεφάλαιο 3, το μειονέκτημα της είναι όμως οτι δεν είναι κατασκευασμένη για να αντιμετωπίζει ανομοιογενή προβλήματα τόσο μεταξύ τους όσο και μεταξύ των κλάσεων του καθενός ξεχωριστά.

4.5 Προσαρμοσμένη Μετα-μάθηση

Σε αυτή την παράγραφο θα γίνει αναφορά σε έναν από τους αλγορίθμους που θα εκπαιδευτούν στη συνέχεια. Ο στόχος να εκπαιδεύσεις ένα μοναδικό μετα-εκπαιδευτή ο οποίος να δουλεύει καλά συγχρόνως για μια κατανομή προβλημάτων είναι πάρα πολύ δύσκολο να επιτευχθεί ειδικά όταν τα προβλήματα διαφέρουν πολύ μεταξύ τους. Μια τέτοια στόχευση μπορεί να οδηγήσει το μοντέλο να βρει κάποιες βέλτιστες λύσεις για μερικά από τα προβλήματα της κατανομής και στα υπόλοιπα να απέχει πάρα πολύ από τη βέλτιστη λύση.

Αυτό το πρόβλημα οδήγησε ώστε να αναπτυχθούν πολλές προσεγγίσεις στις οποίες δημιουργήθηκαν μοντέλα προσαρμοσμένης μετα-μάθησης (31). Μια σημαντική πρόταση έγινε από (32). Στην συγκεκριμένη εργασία πρότειναν τη δημιουργία πολλών παραμέτρων που εκπαιδευόνταν για το κάθε πρόβλημα ξεχωριστά και για την κάθε κλάση ξεχωριστά. Μια πρόταση που δουλεύει αρκετά καλά αφού το μοντέλο εξιδεικεύεται και στο κάθε πρόβλημα ξεχωριστά αλλά και στην κάθε κλάση ακόμα και αν ανήκει σε ένα μόνο πρόβλημα. Απλά η συγκεκριμένη διαδικασία χρειάζεται πολλά προβλήματα με λίγα δεδομένα για να μπορεί να γενικεύση καλύτερα (32). Άλλες προτάσεις επικεντρώνονται στη δημιουργία παραμέτρων για το κάθε πρόβλημα ξεχωριστά χωρίς να συμπεριλαμβάνουν την πιθανότητα της ανισορροπίας των κλάσεων σε ένα πρόβλημα (39).

Μια ακόμα ενδιαφέρουσα τεχνική που έχει χρησιμοποιηθεί από (35) είναι να χωρίστουν τα βάρη του δικτύου σε δύο κατηγορίες βαρών, στα βάρη του δικτύου που αφορούν ένα συγκεκριμένο πρόβλημα και στα βάρη του δικτύου που αφορούν όλα τα προβλήματα της διαδικασίας. Μια τέτοια τεχνική έχει τα πλεονεκτήματα της γιατί μπορεί και να εξιδεικεύσει στο κάθε πρόβλημα και στα χαρακτηριστικά του και να βρει τη βέλτιστη λύση για εκείνο αλλά και να γενικεύσει αφού υπάρχουν πολλά βάρη που αφορούν όλη την κατανομή των προβλημάτων.

Απλά σε μια τέτοια διαδικασία δεν είναι σαφές πως μπορεί να αλλάξει ανάλογα με το πρόβλημα η σημασία της συνολικής μάθησης σε σχέση με τη μάθηση μόνο από τα δεδομένα του προβλήματος. Οπότε αντιμετωπίζεται η πληροφορία που παίρνει το κάθε πρόβλημα από τη μετα-μάθηση και από την εξιδεικευμένη μάθηση για το κάθε πρόβλημα σαν δύο συνισταμένες που θα συνεισφέρουν για το κάθε πρόβλημα με τον ίδιο τρόπο. Αυτή είναι και η βασική διαφορά με την οπτική της Προσαρμοσμένης Μετα-Μάθησης που θα χρησιμοποιηθεί που συνεχίζει να θεωρεί αυτές τις δύο πηγές πληροφορίας (μετα-μάθηση, εξιδεικευμένη μάθηση) ως τις συνισταμένες της εκπαίδευσης για το κάθε πρόβλημα, αλλά μπορεί να διαφέρει η επίδραση τους ανάλογα με το πρόβλημα που είναι προς αντιμετώπιση.

Μια ακόμα πιο κοντινή οπτική σε αυτή που θα παρουσιαστεί στην συνέχεια εμφανίζεται στην εργασία του (15). Στη συγκερικένη εργασία αναφέρονται σε μοντέλα μετα-

μάθησης σε διαφορετικά πεδία και παρομοιάζουν τις αλλαγές των βημάτων κλίσης της κάθε παραμέτρου στο κάθε πρόβλημα ως μια εκ των υστέρων κατανομή συμπερασμάτων για την κάθε παράμετρο ξεχωριστά. Με αυτή την παρομοίωση, θεωρείται η συμπεριφορά μιας παραμέτρου ανάλογα με τα βήματα κλίσης της, οτι συμπεριφέραι σαν τυχαία μεταβλητή σε ένα Μπευζιανό περιβάλλον (15).

Σε ένα στοχαστικό περιβάλλον αναφέρεται και η δημοσίευση του (38). Η συγκεκριμένη δουλειά αντιμετωπίζει τις παραμέτρους του προβλήματος σαν τυχαίες μεταβλητές χωρίς να εισάγει ένα Μπευζιανό περιβάλλον. Οι παράμετροι που είναι τυχαίες μεταβλητές είναι εκείνες που αφορούν το κάθε υποπρόβλημα της κατανομής ξεχωριστά και όχι παραμέτρους που είναι κοινές για όλα τα προβλήματα της διαδικασίας. Όλες αυτές οι προσπάθειες έχουν ως στόχο να αντιμετωπίσουν την αβεβαιότητα που δημιουργείται κατά την εκκίνηση ενός προβλήματος με λίγα δεδομένα.

Αρχικά ο στόχος μας σε αυτή την εργασία είναι να αντιμετωπιστούν τα προβλήματα της ανισσοροπίας των κλάσεων, των προβλημάτων και των προβλημάτων που μπορεί να περιέχουν κλάσεις οι οποίες δεν υπήρχαν εώς τώρα. Αυτή η αντιμετώπιση καθιστά απαραίτητη τη χρήση της προσαρμοσμένης μετα-μάθησης. Η προσαρμοσμένη μετα-μάθηση (Task Adaptive Meta Learning) βασίζεται όπως αναφέρθηκε και προηγουμένως στη μετα-μάθηση σε διαφορετικά πεδία, οπότε ακολουθεί την αρχιτεκτονική τους αλλά θα χρησιμοποιηθούν τρεις διαφορετικές παραμέτρους ω^τ , γ^τ , z^τ για να λυθούν τα προβλήματα που αναφέρθηκαν παραπάνω.

Η πρώτη τυχαία μεταβλητή που θα χρησιμοποιηθεί είναι η ω^τ για να αντιμετωπιστεί το πρόβλημα της ανισσοροπίας των κλάσεων. Η πρώτη αλλαγή που πρέπει να γίνει αφορά το ποσοστό εκμάθησης (learning rate) για το βήμα κλίσης της κάθε κλάσης του προβλήματος σε κάθε βήμα βελτιστοποίησης όλης της διαδικασίας. Πιο συγκεκριμένα, για κάθε κλάση $c = 1, ..., C$, δημιουργείται και ένα συγκεκριμένο διάνυσμα $\omega^\tau = (\omega_1^\tau, ..., \omega_C^\tau) \in [0, 1]^C$ τα οποία θα πολλαπλασιάζονται με το συγκεκριμένο βήμα κλίσης $\nabla_\theta \mathcal{L}(\theta, \mathcal{D}^\rfloor)$ για την κάθε κλάση όπου $\mathcal{D}^c$ είναι ο αριθμός παρατηρήσεων για την κλάση c. Η εξαγωγή των παραμέτρων ω_c^τ έχει γίνει μέσω μίας softmax συνάρτησης και αναμένεται ότι θα είναι μεγάλες για τις μικρές κλάσεις του προβλήματός έτσι ώστε να προσμετρόνται περισσότερο σε κάθε βήμα σύγκλισης των παραμέτρων.

Το αποτέλεσμα αυτό θα έχει να έρχονται σε μια ισορροπία οι κλάσεις του προβλήματος, ανεξαρτήτως των δεδομένων που υπάρχουν αρχικά. Συνεπώς, η παράμετρος ω^τ θα είναι διαφορετική για την κάθε κλάση ενός προβλήματος και θα αφορά την αρχική διαδικασία εκπαίδευσης του κάθε προβλήματος ξεχωριστά. Άρα, αν μια κλάση υπάρχει σε παραπάνω προβλήματα η τιμή του ω^τ θα έχει να κάνει μόνο με τον αριθμό των παρατηρήσεων που εκπροσωπούνται από την συγκεκριμένη κλάση στο συγκεκριμένο πρόβλημα. Έτσι η ίδια κλάση σε διαφορετικά προβλήματα θα έχει και διαφορετικές τιμές του ω^τ.

Στο επόμενο επίπεδο της μάθησης θα εισαχθεί τυχαία μεταβλητή για το δεύτερο πρόβλημα που πρέπει να αντιμετωπιστεί, που είναι η ανισσοροπία δεδομένων μεταξύ των προβλημάτων. Για να ελεχθεί αυτό το πρόβλημα θα δημιουργηθεί για το κάθε πρόβλημα της διαδικασίας ένας διαφορετικός πολλαπλασιαστής $\gamma^\tau = (\gamma_1^\tau, ..., \gamma_C^\tau) \in [0, \infty)^L$, όπου για το κάθε επίπεδο του δικτύου ισχύει $l = 1, ..., L$ έτσι ώστε το ποσοστό εκμάθησης να μετατρέπεται σε $\gamma_1^\tau \alpha, ..., \gamma_L^\tau \alpha$. Ο πολλαπλασιαστής γ^τ θα να είναι μεγάλος για τα μεγάλα προβλήματα της διαδικασίας και να είναι μικρότερος για τα μικρά προβλήματα της διαδικασίας.

Αυτή η προσθήκη θα βοηθήσει έτσι ώστε τα μικρότερα προβλήματα να επηρεάζονται περισσότερο από τη μετα-μάθηση της διαδικασίας σε κάθε επίπεδο του δικτύου αφού οι αλλαγές τους θα επηρεάζονται περισσότερο από το βήμα κλίσης α . Για να μειωθεί η διαφορά μεταξύ μεγάλων και μικρών προβλημάτων της διαδικασίας θα οριστεί ο συγκεκριμένος πολλαπλασιαστής γ^τ μέσω μιας εκθετικής συνάρτησης. Οπότε αυτός ο πολλαπλασιαστής θα επιτρέπει στα μεγάλα προβλήματα να φεύγουν πιο γρήγορα από την αρχική τιμή της παραμέτρτου ϑ και έτσι να χρησιμοποιείτε παραπάνω η πληροφορία τους για την εκπαίδευση τους.

Για να αντιμετωπισττεί το πρόβλημα το ότι μερικές κλάσεις δεν υπάρχουν σε κάποια προβλήματα θα χρησιμοποιηθεί η παρακάτω διαδικασία. Κατασκευάζεται μία παράμετρο z^τ που διαμορφώνει την αρχική παράμετρο θ για το κάθε πρόβλημα. Το z^τ κατασκευάζεται έτσι ώστε να μπορεί να αλλάξει το αρχικό σημείο εκκίνησης για το θ γιατί τα προβλήματα που δεν εμπεριέχουν καθόλου κάποιες κλάσεις μπορούν να επηρεαστούν πάρα πολύ από την αρχική θέση του θ η οποία θα ήταν ίδια για όλα τα προβλήματα. Πιο συγκεκριμένα θα χρησιμοποιηθεί για τα βάρη του συννελικτικού δικτύου $z^\tau = 1 + \tilde{z^\tau}$ και για τα biases $z^\tau = \tilde{z^\tau}$,τα οποία αλλάζουν την παράμετρο ϑ κατά τον ακόλουθο τρόπο: $\theta_0 \leftarrow \theta \circ z^\tau$ για τα βάρη και $\theta_0 \leftarrow \theta + z^\tau$ για την κλίση όπου το θ_0 συμβολίζει την νέα τιμή της παραμέτρου μετά από τη διόρθωση του z^τ. Θα συμβολίζεται αυτή η πράξη ως $\theta_0 \leftarrow \theta \circ z^\tau$ (31), (40).

Άρα συνοψίζοντας, βάζοντας και τις τρεις αυτές παραμέτρους μπορούν να γραφτούν οι παρακάτω ορισμοί για να αναπτυχθεί το πλαίσιο στο οποίο στηρίζονται οι αλλαγές, για να καλυφθούν τα τρία προβλήματα ισορροπίας των δεδομένων σε ένα πρόβλημα:

$\theta_0 \leftarrow \theta \circ z^\tau$

$\theta_k = \theta_{k-1} - \gamma^\tau \circ \alpha \circ \sum \nabla_\theta L(\theta_{k-1}, \mathcal{D}^\rfloor)$

$k = 1, ..., K$

όπου το α είναι ο πολλαπλασιαστής για το βήμα κλίσης για κάθε διαφορετικό πρόβλημα (41) της διαδικασίας και το θ_k είναι διαφορετικό για το κάθε πρόβλημα της διαδικασίας.

4.6 Μπεϋζιανή Προσαρμοσμένη Μετα-μάθηση

Έγινε αναφορά σε αλγορίθμους προσαρμοσμένης μετα-μάθησης και ορίστηκε ο τρόπος λειτουργίας τους που βασίζεται στις τρεις παραμέτρους που χρειάζονται για να ισορροπήσουν τα δεδομένα του προβλήματος. Το επόμενο βήμα είναι να γίνει ανάλυση για τους αλγορίθμους της μπεϋζιανής προσαρμοσμένης μετα-μάθησης. Η κύρια διαφορά με τους προηγούμενους αλγορίθμους, είναι ότι θα γίνει ο ορισμός τριών καινούργιων παραμέτρων $\tilde{\omega_c^\tau}$, $\tilde{\gamma^\tau}$, $\tilde{z^\tau}$ ως τυχαίες μεταβλητές που η κάθε μία από αυτές θα ακολουθεί συγκεκριμένη κατανομή. Επίσης αυτό το δίκτυο θα χρησιμοποιηθεί στα πειράματα που θα αναφερθούν στη συνέχεια.

Στο μπεϋζιανό πλαίσιο είναι πολύ εύκολο να εισαχθεί η τυχαιότητα στις εκ των υστέρων κατανομές αυτών των παραμέτρων. Λόγω αυτής της ευκολίας δίνεται η δυνατότητα να φτιαχτεί πιο συγκροτημένη πληροφορία για αυτές τις παραμέτρους με την τυχαιότητα που μπορεί να εισαχθεί στις εκ των υστέρων κατανομές. Για τον ορισμό αυτών των παραμέτρων σύμφωνα με τους όρους της Μπεϋζιανής Στατιστικής θα ακολουθηθεί η παρακάτω διαδικασία.

Αρχικά θα πρέπει να οριστεί για τα δεφομένα οτι $X^\tau = x_n^T$ και $Y^\tau = y_n^T$ για το σύνολο εκπαίδευσης όπου $n = 1, ..., N_\tau$, και $X^\tau = x_m^T$ και $Y^\tau = y_m^T$ για το σύνολο που θα ελεγχθούν τα δεδομένα όπου $m = 1, ..., M_\tau$. Θα οριστεί ώς φ^τ η συλλογή των τριών παραμέτρων $\tilde{\omega_c^\tau}$, $\tilde{\gamma^\tau}$, $\tilde{z^\tau}$. Για το κάθε πρόβλημα της διαδικασίας Θα ακολουθηθεί ο παρακάτω ορισμός για τον προσδιορισμό των τριών παραμέτρων που έχουν δημιουργηθεί (15), (38):

$$p(\tilde{Y^\tau}, \tilde{Y^\tau}, \tilde{\varphi^\tau} | \tilde{X^\tau}, \tilde{X^\tau}; \theta) = p(\varphi^\tau) \prod_{i=1}^{N_\tau} p(y_n^\tau, | x_n^\tau, \varphi^\tau; \theta) \prod_{i=1}^{M_\tau} p(y_m^\tau, | x_m^\tau, \varphi^\tau; \theta)$$

Αρχικά, ο παραπάνω τύπος αφορά το ίδιο θ για όλα τα προβλήματα της διαδικασίας. Οπότε, η Μπεϋζιανή Προσαρμοσμένη μετα-Μάθηση που θα χρησιμοποιηθεί είναι βασισμένη στη μετα-μάθηση σε διαφορετικά πεδία συνδυασμένη με τις τρεις παραμέτρους που θα δημιουργηθούν για να μπορέσουν να αντιμετωπιστούν προβλήματα ανισορροπίας δεδομένων. Αυτές οι παράμετροι θα αντιμετωπιστούν ως τυχαίες μεταβλητές σε ένα Μπεϋζιανό Περιβάλλον. Στην συνέχεια θα γίνει αναφορά στο πως θα δημιουργηθούν αυτές οι τρεις παράμετροι και σε τι στοιχεία του σετ δεδομένων θα βασιστούν για να παραχθούν. Ακόμα σημαντικό ρολο παίζει και η σειρά με την οποία θα δημιουργηθούν. Αφού για παράδειγμα η παράμετρος για την ανισορροπία των κλάσεων σε κάθε πρόβλημα ξεχωριστά θα δημιουργηθεί σε προγενέστερο χρόνο από οτι οι άλλες δύο παράμετροι.

4.7 Το Στοχαστικό πλαίσιο των τριών παραμέτρν

Στην προηγούμενη ενότητα παρουσιάστηκε ο τύπος με τον οποίο θα υπολογίζονται οι τρεις στοχαστικές παράμετροι της διαδικασίας. Ο στόχος της μάθησης για το κάθε πρόβλημα τ της διαδικασίας είναι η μεγιστοποίηση της δεσμευμένης λογαριθμικής πιθανοφάνειας, για τα δεδομένα εκπαίδευσης D_τ και για τα δεδομένα ελέγχου $\tilde{D}_\tau$: $p(\tilde{Y}^\tau, \tilde{Y}^\tau | \tilde{X}^\tau, \tilde{X}^\tau; \theta)$. Ο υπολογισμός αυτός θα δώσει την κατάλληλη τιμή για την κάθε μια από τις παραμέτρους του προβλήματος για την καλύτερη απόδοση του μοντέλου. Οι τρεις παράμετροι που έχουν δημιουργηθεί $\tilde{\omega^\tau}$, $\tilde{\gamma^\tau}$, $\tilde{z^\tau}$ θα υπολογίζονται στο ίδιο δίκτυο για να μειώνεται το υπολογιστικό κόστος.

Στην συνέχεια, σύμφωνα με την (27), αφαιρείται η συσχέτιση που υπάρχει στο σύνολο ελέγχου $\tilde{D}_\tau$ για την εκ των υστέρων κατανομή, έτσι ώστε οι δύο διαδικασίες να διαφέρουν σαφώς. Η μία είναι για τη διαδικασία της μετα-μάθησης που αφορά σε όλο το σύνολο των δεδομένων εκπαίδευσης και η άλλη που αφορά το μετα-έλεγχο όπου δεν είναι γνωστές οι τιμές των κλάσεων των παρατηρήσεων των δεδομένων ελέγχου. Η τελική μορφή της εκ των υστέρων κατανομής είναι $q(\varphi^\tau | D^\tau; \psi)$, όπου για την κάθε μια από τις τρεις στοχαστικές παραμέτρους η κατανομή τους μπορεί να είναι διαφορετική.

Οπως αναφέρθηκε και στις προηγούμενες ενότητες σημαντικό ρόλο στη μετα-μάθηση σε διαφορετικά πεδία διαδραματίζει και το βήμα κλίσης. Το βήμα κλίσης είναι υπεύθυνο για την απομάκρυνση της παραμέτρου από την αρχική της τιμή. Στη Μπεϋζιανή προσαρμοσμένη μετα-μάθηση που θα χρησιμοποιηθεί ο (25) έχει ορίσει το βήμα κλίσης. Θα παρατεθεί το κατώτερο όριο αλλαγής του βήματος κλίσης που χρειάζεται για τη διαδικασία μετα-μάθησης του προβλήματος:

$$L^\tau_{\theta,\psi} = \frac{N_\tau + M_\tau}{M_\tau} \sum_{m=1}^{M_\tau} E_{q(\varphi^\tau | D^\tau; \psi)} [\log p(\tilde{y}^\tau_m | \tilde{x}^\tau_m, \varphi^\tau_s; \theta)] - KL[q(\varphi^\tau | D^\tau; \psi) | p(\varphi^\tau)]$$

.

Γίνεται η υπόθεση, οτι η εκ των υστέρων κατανομή $q(\varphi^\tau | D^\tau; \psi)$ μπορεί να απεικονίσει όλες τις παραμέτρους του προβλήματος και για όλες τις διαστάσεις τους:

$$q(\varphi^\tau | D^\tau; \psi) = p(\varphi^\tau) \prod_c q(\tilde{\omega^\tau_c}, | D^\tau; \psi) \prod_l q(\tilde{\gamma^\tau_l}, | D^\tau; \psi) \prod_i q(\tilde{z^\tau_i}, | D^\tau; \psi)$$

.

Για την κάθε κλάση του προβλήματος η εκ των υστέρων κατανομή $q(\varphi^\tau | D^\tau; \psi)$ ακολουθεί Γκαουσιανή κανονική κατανομή με εκπαιδευμένη μέση τιμή και διασπορά. Ακόμα η εκ των προτέρων κατανομή $p(\varphi^\tau)$ ακολουθεί κανονική κατανομή $N(0, 1)$ και η KL απόκλιση έχει την ακόλουθη μορφή από την (28). Η τελική μορφή της πληροφορίας της μετα-μάθησης μετά από διαδιακασία Monte-Carlo (MC) είναι η ακόλουθη:

$$\min_{\theta,\psi} \frac{1}{M_\tau} \sum_{m=1}^{M_\tau} \frac{1}{S} \sum_{i=1}^{S} -\log p(\tilde{y}_m^\tau | \tilde{x}_m^\tau, \varphi_s^\tau; \theta) + \frac{1}{N_\tau + M_\tau} KL[q(\varphi^\tau | D^\tau; \psi) | p(\varphi^\tau)]$$

.

όπου $\varphi_s^\tau \approx q(\varphi^\tau | D^\tau; \psi)$, αυτος ορισμός της φ^τ βοηθάει στο να έχουμε μια πιο σαφή εκτίμηση για τη συνάρτηση (28).

$$p(\tilde{y}_*^\tau | \tilde{x}_*^\tau; \theta) = E_q[p(\tilde{y}_*^\tau | \tilde{x}_*^\tau, \varphi^\tau; \theta)] \approx \frac{1}{S} \sum_{i=1}^{S} p(\tilde{y}_*^\tau | \tilde{x}_*^\tau, \varphi_s^\tau; \theta)$$

.

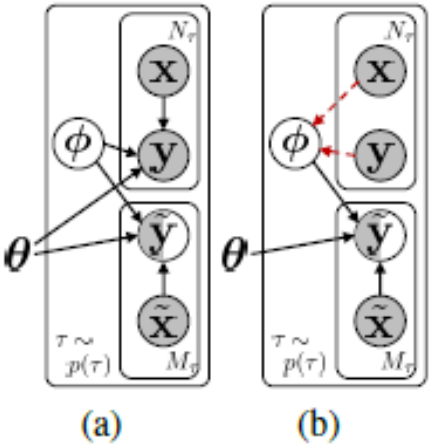
$$\varphi_s^\tau \approx q(\varphi^\tau | D^\tau; \psi)$$

.

Μια εναλλακτική μορφή της μέσης τιμής για να είναι πιο εύκολος ο υπολογισμός της θα παρατεθεί στην επόμενη ισότητα, αφού θα χρειάζεται να υπολοζίσουμε τη μέση τιμή μόνο της συνάρτησης φ^τ.

$$E_q[p(\tilde{y}_*^\tau | \tilde{x}_*^\tau, \varphi^\tau; \theta)] \approx p(\tilde{y}_*^\tau | \tilde{x}_*^\tau, E_q[\varphi^\tau]; \theta)$$

.



Σχήμα 13: μετα-εκπαίδευση, μετα-έλεγχος

4.8 Κωδικοποίση των Δεδομένων

Στο συγκεκριμένο πρόβλημα θα συμμετέχουν πολλά σετ δεδομένων με διαφορετική πληροφορία το κάθε ένα. Ο βασικός στόχος από τη στιγμή που φτιάχνεται στο μοντέλο μια καινούργια κατανομή $q(\varphi^\tau|D^\tau;\psi)$ για κάποιες παραμέτρους του είναι ο τρόπος που θα οριστεί το σετ εκπαίδευσης των δεδομένων έτσι ώστε η πληροφορία που θα παίρνει η κατανομή της κάθε παραμέτρου να μην είναι τετριμμένη (**?**). Γεγονός που θα οδηγούσε ένα μοντέλο Μπεϋζιανής προσαρμοσμένης μετα-μάθησης να μην έχει ουσία αφού οι παράμετροι που ορίζει σαν τυχαίες μεταβλητές δεν θα άλλαζαν καθόλου και θα επανέρχοταν σε ένα μοντέλο απλώς προσαρμοσμένης μετα-μάθησης.

Αρχικά, για να επιτευχθεί αυτός ο σκοπός πρέπει η διαδικασία μάθησης να περιλαμβάνει όλη την απαιτούμενη πληροφορία από το σετ δεδομένων. Συνοψίζοντας τις τρεις βασικές διαφορές που πρέπει να αντιμετωπιστούν στο πρόβλημα που έχει τεθεί : ανισορροπία κλάσεων, ανισορροπία προβλημάτων, κλάσεις που δεν υπάρχουν στην κατανομή των προβλημάτων. Πολύ γνωστές τεχνικές που χρησιμοποιούνται είναι το άθροισμα (Sum-Pooling) (36) και ο μέσος όρος (Mean-Pooling) (37). Στην πρώτη περίπτωση, το άθροισμα (Sum-Pooling) χρησιμοποιοείται σαν ένας κωδικοποιητής του συνόλου, όπου το κάθε παράδειγμα του σετ δεδομένων αθροίζεται ανάλογα με το φίλτρο που χρησιμοποιείται και δίνεται σαν αποτέλεσμα το άθροισμα αυτών των στοιχείων για να παραχθεί ένα καινούργιο σετ δεδομένων που θα εμπεριέχει όλη την αρχική πληροφορία σε πιο συμπηκνωμένη μορφή. Στην περίπτωση του (Mean-Pooling) γίνεται ακριβώς η ίδια διαδικασία κωδικοποίησης, μόνο που σε αυτή την περίπτωση, από τα δεδομένα που χρησιμοποιεί ο κωδικοποιητής προκύπτει σαν αποτέλεσμα ο μέσος όρος τους, ανάλογα και πάλι με τα φίλτρα που έχουν χρησιμοποιηθεί. Συμπερασματικά, από το άθροισμα (Sum-Pooling) προκύπτουν νέα στοιχεία που περιέχουν το άθροισμα των χαρακτηριστικών της εικόνας που χρησιμοποιούνται ενώ στο μέσο όρο (Mean-Pooling) τα νέα στοιχεία χρησιμοποιούν το μέσο όρο. Άρα τα δεδομένα του προβλήματος δέχονται κάποιους μετασχηματισμούς έτσι ώστε ο κωδικοποιητής να αποκτήσει περισσότερη πληροφορία.

Οι συγκεκριμένες μέθοδοι ενώ χρησιμοποιούνται σε πάρα πολλές εφαρμογές (36), (37) στην συγκεκριμένη περίπτωση δεν μπορεί να είναι πολύ χρήσιμες. Στη διαδικασία που θα ακολουθηθεί τα δεδομένα του προβλήματος ανήκουν σε ένα σύνολο από σύνολα κλάσεων, αφού υπάρχουν πολλά διαφορετικά προβλήματα με πολλές διαφορετικές κλάσεις που κάποιες μπορεί να εμφανίζονται μόνο σε ένα πρόβλημα και άλλες σε όλα τα προβλήματα τις διαδικασίας.

Σε κατανομές προβλημάτων που ύπαρχει διαφορά στα παραδείγματα της κάθε κλάσης του κάθε προβλήματος και στην συνέχεια διαφορές μεταξύ των προβλημάτων, χρειάζεται μια διαδικασία η οποία πρώτα θα ασχολείται εσωτερικά με το κάθε πρόβλημα της κατανομής και θα προσπαθεί να επιδρά στην ανισορροπία των κλάσεων και στην

συνέχεια θα αντιμετωπίζει την ανισορροπία που υπάρχει μεταξύ των προβλημάτων. Ένας ακόμα περιορισμός που υπάρχει στο μέσο όρο (Mean-Pooling) είναι, το τι στοιχεία χρησιμοποιεί για να περιγράψει ένα πρόβλημα, αφού δεν μπορεί να αναγνωρίσει τον αριθμό των στοιχείων σε κάθε πρόβλημα. Αυτή η αδυναμία μπορεί να το οδηγήσει σε πολύ χαμηλότερες επιδόσεις σε περιπτώσεις που ο αριθμός των δεδομένων σε κάθε πρόβλημα της διαδικασίας μπορεί να διαφέρει. Οπότε ο κωδικοποιητής που πρέπει να χρησιμοποιηθεί στο συγκεκριμένο πρόβλημα πρέπει να μπορεί να δώσει ιεραρχικά την πληροφορία αρχικά για το κάθε πρόβλημα ξεχωριστά και στη συνέχεια για το κάθε πρόβλημα σε σχέση με τα άλλα προβλήματα.

Σύμφωνα με αυτούς τους περιορισμούς στην συγκεκριμένη εργασία θα χρησιμοποιηθεί ένας κωδικοποιητής δύο επιπέδων για τα δεδομένα. Ο συγκεκριμένος κωδικοποιητής πρώτα κωδικοποιεί το κάθε πρόβλημα ξεχωριστά σαν ένα σύνολο και στη συνέχεια στο δεύτερο επίπεδο βλέπει τη διαδικασία ως ένα σύνολο προβλημάτων και προσπαθεί εκεί να αντιμετωπίσει τις ανισσοροπίες που υπάρχουν μεταξύ των προβλημάτων. Η πρώτη πληροφορία που εισάγεται για το κάθε σετ δεομένων του προβλήματος στον κωδικοποιητή είναι ο αριθμός των στοιχείων της κάθε κλάσης του κάθε προβλήματος. Αυτός ο αριθμός όμως δεν είναι ικανός για να περιγράψει από μόνος του την κατανομή ενός προβλήματος. Για παράδειγμα μπορεί να υπάρχει μια κλάση η οποία να έχει πάρα πολλές εγγραφές αλλά όλες αυτές οι εγγραφές να μην διαφέρουν σχεδόν καθόλου ή ακόμα και να είναι ακριβώς οι ίδιες εγγραφές.

Σε μια τέτοια περίπτωση θα υπήρχε από τον κωδικοποιητή μια υπερεκτίμηση της πληροφορίας, αφού θα βασιζόταν στον αριθμό των δεδομένων της κλάσης ή σε αντίθετη περίπτωση με λίγα δεδομένα που θα διέφεραν αρκετά μεταξύ τους θα υπήρχε μια υποεκτίμηση της πληροφορίας. Συνεπώς, για να χρησιμοποιήσει ο κωδικοποιητής ακόμα περισσότερη πληροφορία για το σετ δεδομένων θα εισαχθεί η μέση τιμή και η διασπορά των δεδομένων για την κάθε κλάση του κάθε προβλήματος. Στο πρώτο επίπεδο κωδικοποίησης αυτά αφορούν την κάθε κλάση του προβλήματος και γίνεται μια κωδικοποίηση ως προς τις κλάσεις του κάθε προβλήματος, όπου αυτή η κωδικοποίηση βοηθά στην δημιουργία της παραμέτρου $\tilde{\omega^{\tau}}$ που είναι υπεύθυνη για το μέγεθος της κάθε κλάσης του κάθε προβλήματος και βοηθάει στο πρόβλημα της ανισορροπίας των κλάσεων σε ένα πρόβλημα.

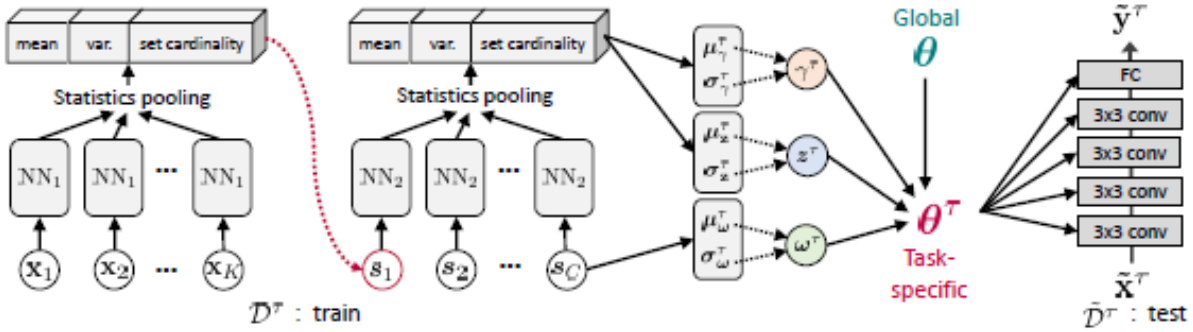
Στο δεύτερο επίπεδο κωδικοποίησης ακολουθείται μια παρόμοια διαδικασία για το σύνολο του συνόλου των προβλημάτων, όπου εδώ το κάθε πρόβλημα αντιμετωπίζεται ακριβώς όπως η κάθε κλάση στο προηγούμενο επίπεδο. Αρχικά, παίρνουμε τον αριθμό των παρατηρήσεων, τη μέση τιμή και την διασπορά του κάθε προβλήματος ξεχωριστά. Ο κωδικοποιητής χρησιμοποιεί αυτή την πληροφορία για να εξάγει συμπεράσματα για το κάθε πρόβλημα σε σχέση με τα άλλα προβλήματα και να μπορεί να δημιουργήσει τις παραμέτρους $\tilde{\gamma^{\tau}}$, $\tilde{z^{\tau}}$.

Βασιμένοι σε αυτή την ιδέα οι (25), ορίσαν την παρακάτω διαδικασία η οποία απότελείται από δύο επίπεδα κωδικοποίησης όπως αναφέρθηκε και νωρίτερα. Ονομάσαν αυτή την διαδικασία στατιστική μελέτη (Statistics Pooling) η οποία χρησιμοποιεί τα στατιστικά στοιχεία του αριθμού των δεδομένων, της μέσης τιμής και της διασποράς όπως φαίνεται παρακάτω και στα δύο επίπεδα:

$$v^\tau = StatisticsPooling(NN_2(s_c)) \ , \ s_c = StatisticsPooling(NN_1(x)_{x \in X_c^\tau})$$

για τις κλάσεις του προβλήματος $c = 1, ..., C$. X_c^τ είναι τα στοιχεία της κλάσης c στο πρόβλημα τ. Τα δίκτυα NN_1, NN_2 είναι νευρωνικά δίκτυα που έχουν παραμετροποιηθεί από το ψ . Το διάνυσμα v^τ εν τέλει περικλύει όλη την πληροφορία για τα δεδομένα του αντίστοιχου προβλήματος. Επίσης οι τυχαίες μεταβλητές $\omega_1^\tau, ..., \omega_C^\tau$ δημιουργούνται από τις $s_1, ..., s_C$ και οι τυχαίες μεταβλητές γ^τ, z^τ που αφορούν το κάθε πρόβλημα ξεχωριστά δημιουργούνται από το v^τ.

;



Σχήμα 14: Κωδικοποίση Δεδομένων

5 Πειράματα

5.1 Εισαγωγή

Στόχος της συγκεκριμένης εργασίας είναι η προσπάθεια αντιμετώπισης ρεαλιστικών προβλημάτων χρησιμοποιώντας τεχνικές βασισμένες στη μετα-μάθηση. Βασιζόμενοι στην εργασία του (25) θα αναπτυχθούν τεχνικές μετα-μάθησης σε διαφορετικά πεδία, τεχνικές μετα-μάθησης με μεταβαλλόμενο ποσοστό μάθησης (learning rate) και η μέθοδος της μπεϋζιανής προσαρμοσμένης μετα-μάθησης. Θα αναπτυχθούν μοντέλα εκπαίδευσης και θα ελεχθούν από δεδομένα ελέγχου. Το σύνολο των δεδομένων αποτελείται από εννέα διαφορετικά σετ δεδομένων με πολλές διαφορέτικες κλάσεις, για να μπορεί να αναπαρασταθεί η ετερόκλητη πληροφορία που χρειάζεται σε ένα τέτοιο πρόβλημα. Θα ακολουθήθει η διαδικασία που χρησιμοποιήθηκε από τον (25).

Αρχικά η μελέτη χωρίζεται σε τρία διαφορετικά πειράματα. Κάθε πείραμα θα περιλαμβάνει διαφορετικά σετ δεδομένων. Στα πρώτα δύο πειράματα θα υπάρχουν δυο σετ δεδομένων ένα κυρίως που θα χρησιμοποιηθεί για το σετ εκπαίδευσης και ένα που θα χρησιμοποιηθεί περισσότερο για το σετ ελέγχου. Στο τρίτο πείραμα θα υπάρχει ένα πρόβλημα με πολλά σετ δεδομένων, πέντε τον αριθμό, εκ των οποίων τα τρία χρησιμοποιούνται σαν σετ εκπαίδευσης.

Τα μοντέλα που θα αναπτυχθούν και στα τρία πειράματα βασίζονται στην ίδια αρχιτεκτονική. Η αρχιτεκτονική και των τριών μοντέλων είναι βασισμένη στη διαδικασία που ακολουθεί ο (15). Το πρώτο μοντέλο που είναι η μετα-μάθηση σε διαφορετικά πεδία ακολουθεί ακριβώς τις ίδιες τεχνικές. Το επόμενο μοντέλο έχει να κάνει με τις αλλαγές που μπορούν να γίνουν στο ποσοστό μάθησης κάτα την διαδικασία εκπαίδευσης των παραμέτρων. Το τρίτο κάνει χρήση της μπεϋζιανής προσαρμοσμένης μετα-μάθησης.

5.2 Μοντέλα που θα συγκριθούν

Οι τεχνικές που θα ακολουθήθουν στην συγκεκριμένη εργασία είναι τρεις. Η πρώτη τεχνική την οποία θα χρησιμοποιηθεί και στα τρία παραδείγματα είναι η μετα-μάθηση σε διαφορετικά πεδία. Θα χρησιμοποιηθεί η μετα-μάθηση σε διαφορετικά πεδία σύμφωνα με την προσέγγιση του (15).

Στη μετα-μάθηση σε διαφορετικά πεδία έχει γίνει εκτενή αναφορά σε προηγούμενα κεφάλαια και για την ικανότητα της να μπορεί να αντιμετωπίζει κατανομές προβλημάτων με πολύ διαφορετική πληροφορία μεταξύ των σετ δεδομένων ή και μεταξύ των κλάσεων ενός σετ δεδομένων. Στα πρώτα δύο πειράματα το μοντέλο θα εκπαιδευτεί σε ένα σετ δεδομένων που θα έχει πολλές διαφορετικές κλάσεις και στο τρίτο πείραμα θα εκπαιδευτεί σε τρία διαφορετικά σετ δεδομένων. Αυτό που θα αξιολογηθεί είναι η ικανότητα ενός τέτοιου μοντέλου να μπορεί μέσω των παραμέτρων του να αναπαριστά

την πληροφορία σε πολύ διαφορετικές κλάσεις στα πρώτα δύο πειράματα και σε πολύ διαφορετικά προβλήματα στο τρίτο πείραμα.

Η μετα-μάθηση με μεταβαλλόμενο ποσοστό μάθησης είναι η δεύτερη τεχνική που θα ακολουθηθεί. Η αρχιτετκτονική της είναι παρόμοια με την μετα-μαθηση σε διαφορετικά πεδία. Η κύρια διαφορά τους εντοπίζεται στο ποσοστό μάθησης. Στην συγκεκριμένη μέθοδο το ποσοστό μάθησης μπορεί να διαφέρει μεταξύ των προβλημάτων της κατανομής. Συνεπώς, θα γίνει η αντίστοιχη παρατήρηση στα αποτελέσματα των πειραμάτων αν η συγκεκριμένη δυνατότητα παίζει ρόλο στην καλύτερη απόδοση του μοντέλου.

Η τρίτη μέθοδος είναι η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση. Η βασική διαδικασία που ακολουθεί το μοντέλο είναι η ίδια με την μετα-μάθηση σε διαφορετικά πεδία. Οι διαφορές εντοπίζονται, αρχικά, στο ποσοστό μάθησης το οποίο μπορεί να μεταβάλλεται ανάλογα με το πρόβλημα. Η δεύτερη διαφορά που είναι το κύριο αντικείμενο της εργασίας, είναι η ύπαρξη τριών παραμέτρων που αντιμετωπίζονται ως τυχαίες μεταβλητές.

Οι συγκεκριμένες παράμετροι αλλάζουν τον τρόπο που κατασκευάζονται οι υπόλοιπες παράμετροι του προβλήματος. Αυτή η αλλαγή γίνεται βασιζόμενη σε τρεις συνιστώτες. Αρχικά, στο μέγεθος της κάθε κλάσης του κάθε προβλήματος. Στην συνέχεια, στην διαφορά των δεδομένων του κάθε προβλήματος σε σχέση με τα άλλα προβλήματα της διαδικασίας. Η τρίτη συνιστώσα που επιδρά στην συγκεκριμένη αλλαγή είναι η ύπαρξη μιας κλάσης μόνο σε ένα πρόβλημα από όλα τα προβλήματα της διαδικασίας.

Στα κεφάλαια των πειραμάτων θα γίνει εκτενής αναφορά στις τιμές που επιλέχθηκαν για την κάθε παράμετρο, στον αριθμό των δεδομένων που εκπαιδεύτηκαν, καθώς και στον αριθμό των δεδομένων που υπάρχουν στα σετ ελέγχου. Τα δύο πρώτα πειράματα περιέχουν ένα σετ δεδομένων εκπαίδευσης ενώ το τρίτο θα περιέχει τρία διαφορετικά σετ εκπαίδευσης. Συνεπώς, τα πρώτα δύο ανήκουν στην κατηγορία προβλημάτων ενος σετ εκπαίδευσης(Any-shot classification) ενώ το τρίτο στην κατηγορία πολλαπλών σετ εκπαίδευσης(Multi-dataset classification)

5.3 Σύνολα Δεδομένων

Τα Σύνολα Δεδομένων που θα χρησιμοποιηθούν αφορούν πολύ διαφορετικά πεδία. Στο σύνολο τους είναι εννέα διαφορετικά σετ δεδομένων εκ των οποίων τα πέντε θα χρησιμοποιηθούν και για την εκπαίδευση του μοντέλου ενώ τα υπόλοιπα τέσσερα θα χρησιμοποιηθούν μόνο στα σετ ελέγχου.

Στα τρία πειράματα που θα πραγματοποιηθούν θα χρησιμοποιηθεί κατά σειρά στο πρώτο πείραμα δύο σετ δεδομένων όπου μόνο το ένα θα είναι για να εκπαιδευτούν τα μοντέλα. Στο δεύτερο πείραμα, δύο σετ δεδομένων όπου και πάλι μόνο το ένα σετ δεδομένων θα αφορά την εκπαίδευση του δεύτερου πειράματος. Τέλος, στο τρίτο

πείραμα θα χρησιμοποιηθούν τα υπόλοιπα πέντε σετ δεδομένων εκ των οποίων μόνο τα τρία θα αξιοποιηθούν για την εκπαίδευση του μοντέλου. Για να γίνει ο έλεγχος των αποτελεσμάτων των μοντέλων θα χρησιμοποιηθούν και τα εννέα σετ δεδομένων. Στο πρώτο και το δεύτερο πείραμα και τα δύο σετ δεδομένων που τα αφορούν και στο τρίτο πείραμα και τα πέντε σετ δεδομένων που περιλαμβάνονται στη συγκεκριμένη διαδικασία.

Στο πρώτο από τα τρία πειράματα θα χρησιμοποιηθεί ένα πάρα πολύ γνωστό σετ δεδομένων για τέτοια προβλήματα το CIFAR 100 (42). Το συγκεκριμένο σετ δεδομένων πήρε το όνομα του από το Ινστιτούτο Έρευνας του Καναδά (Canadian Institute for Advanced Research) και αποτελείται από 60000 διαφορετικές εικόνες. Αρχικά, το σετ δεδομένων αποτελείται από 20 υπερκλάσεις όπου η κάθε υπερκλάση εμπεριέχει πέντε κλάσεις. Οπότε στο σύνολο υπάρχουν στο σετ δεδομένων 100 κλάσεις με 600 παρατηρήσεις στην κάθε κλάση. Το αντικείμενο της κλαθε κλάσης είναι μια εικόνα 32*32 pixels. Ο διαχωρισμός των δεδομένων που θα ακολουθηθεί είναι 64 κλάσεις για το σετ εκπαίδευσης, 16 κλάσεις για το σετ επικύρωσης και 20 κλάσεις για το σετ ελέγχου.

Τα δεδομένα αυτού του προβλήματος αν και στο πρώτο πείραμα θα χρησιμοποιηθούν μόνο αυτά για να εκπαιδευτούν τα μοντέλα παρουσιάζουν αρκετό ενδιαφερόν ως προς την πληροφορία που διαθέτουν. Όπως αναφέρθηκε και νωρίτερα οι κλάσεις του σετ δεδομένων είναι 100. Αυτές οι κλάσεις χωρίζονται σε 20 υπερκλάσεις οι οποίες έχουν εικόνες που ανήκουν σε μια οικογένεια αντικειμένων τις περισσότερες φορές. Κάποιες από τις οικογένειες αντικειμένων είναι ψάρια και θαλάσσια θηλαστικά ενώ άλλες είναι ηλεκτρικές συσκευές , οχήματα ή ακόμα φρούτα και λαχανικά. Το γεγονός αυτό έχει σαν αποτέλεσμα από μόνο του το σετ δεδομένων να δίνει την δυνατότητα στο μοντέλο να αποκτά πολύ διαφορετική πληροφορία και μέσα από αυτή να προσπαθεί να εξάγει συμπεράσματα για εικόνες που μπορεί να ανήκουν σε πολύ διαφορετικές κλάσεις. Συνεπώς ακόμα και σε αυτή την περίπτωση που τα μοντέλα θα εκπαιδευτούν μόνο σε ένα πρόβλημα ,η πληροφορία είναι τόσο διαφορετική που καθιστά πολύ δύσκολη την ταξινόμηση τόσο διαφορετικών εικόνων.

Το δεύτερο σετ δεδομένων που θα χρησιμοποιηθεί στο πρώτο πείραμα είναι από (43). Το συγκεκριμένο σετ δεδομένων ονομάζεται SVHN (Street View House Numbers) και χρησιμοποιείται σε πολλές περιπτώσεις σαν σετ ελέγχου για σετ δεδομένων όπως το CIFAR 10 και το CIFAR 100. Τα δεδομένα του είναι 26.032 έγχρωμες εικόνες που απεικονίζουν μονοψήφιους αριθμούς από το 0 εώς στο 9 οι οποίοι απεικονίζονται σε έγχρωμες εικόνες 32*32 pixels. Στην ίδια ακριβώς μορφή με το σετ εκπαίδευσης. Αυτά τα δύο σετ δεδομένων θα περιλαμβάνονται στο πρώτο πείραμα.

Στο δεύτερο πείραμα θα χρησιμοποιηθούν δύο σετ δεδομένων ακριβώς όπως στο πρώτο. Το πρώτο σετ δεδομένων θα είναι ένα υποσύνολο του Imagenet από (9).

Το συγκεκριμένο σετ δεδομένων είναι πολύ γνωστό και χρησιμοποιείται ευρέως σε προβλήματα μάθησης με λίγα δεδομένα. Το αρχικό σετ δεδομένων αποτελείται από 150 χιλιάδες εικόνες οι οποίες ανήκουν σε 1000 διαφορετικές κλάσεις. Αυτες οι κλάσεις μπορεί να διαφέρουν πολύ μεταξύ τους και ως επί το πλείστον αφορούν εικόνες από ζώα ή φυτά.

Το υποσύνολο που θα χρησιμοποιηθεί θα αποτελείται από 60000 παραδείγματα όπου όπως και στο προηγούμενο πείραμα θα παρθούν 100 κλασεις και 600 παραδείγματα για την κάθε κλάση όπως έκανε ο (9). Οι συγκεκριμένες εικόνες έχουν μετασχηματιστεί σε διαστάσεις 84*84. Ο διαχωρισμος για το σετ δεδομένων θα γίνει με τον ίδιο τρόπο ακριβώς όπως και με το CIFAR σετ δεδομένων του πρώτου πειράματος. Ο διαχωρισμός των δεδομένων που θα ακολουθηθεί είναι 64 κλάσεις για το σετ εκπαίδευσης, 16 κλάσεις για το σετ επικύρωσης και 20 κλάσεις για το σετ ελέγχου. Το δεύτερο σετ δεδομένων που θα χρησιμοποιηθεί είναι το σετ δεδομένων CUB. Τα δεδομένα αυτά περιλαμβάνουν 11.788 εικόνες από 200 διαφορετικά είδη πτηνών. Οι εικόνες και του συγκεκριμένου σετ δεδομένων έχουν μετασχηματιστεί για να έχουν διαστάσεις 84*84.

Το τρίτο πείραμα θα αποτελείται από 5 διαφορετικά σετ δεδομένων εκ των οποίων τα τρία θα συνεισφέρουν στην εκπαίδευση και τον έλεγχο των μοντέλων ενώ τα υπόλοιπα δύο θα συνεισφέρουν μόνο στον έλεγχο των δεδομένων. Το πρώτο σετ δεδομένων που θα χρησιμοποιηθεί στην διαδικασία της εκπαίδευσης είναι το Aircraft από (44). Το σετ δεδομένων περιέχει 102 διαφορετικού είδους αεροσκάση και για την κάθε κατηγορία αεροσκάφους 100 διαφορετικές εικόνες. Οι φωτογραφίες έχουν κοπεί κατάλληλα έτσι ώστε να μην περιλαμβάνουν άλλα αεροσκάφη είτε από που προέρχεται η κάθε εικόνα. Ο διαχωρισμός των κλάσεων που θα ακολουθηθεί είναι 70 κλάσεις για το σετ εκπαίδευσης, 15 κλάσεις για το σετ επικύρωσης και 15 κλάσεις για το σετ ελέγχου.

Το επόμενο σετ δεδομένων με το οποίο εκπαιδεύονται τα μοντέλα στο τρίτο πείραμα είναι το Quick Draw. Είναι ένα σύνολο δεδομένων που αποτελείται από 50 εκατομμύρια σκίτσα 345 διαφορετικών κλάσεων. Στο συγκεκριμένο πρόβλημα θα κατανεμηθούν οι κλάσεις σε 241 για τα σετ εκπαίδευσης, 52 κλάσεις για το σετ επικύρωσης και τις άλλες 52 κλασεις για το σετ ελέγχου. Η κάθε μία από τις κλάσεις θα έχει 200 παραδείγματα. Αυτή ακριβώς την διαδικασία για αυτό το υπόσυνολο του σετ δεδομένων ακολούθησε ο (46). Αυτό το σύνολο δεδομένων αν κάποιος παρατηρήσει τις εικόνες είναι ένα πάρα πολύ δύσκολο πρόβλημα ακόμα και αν το μοντέλο εκπαιδευόνταν πάνω μόνο στο συγκεκριμένο πρόβλημα. Αφού ενω διαθέτει πάρα πολλές κατηγορίες, οι εικόνες δεν διαφέρουν αρκετά μεταξύ τους και κάνει ακόμα δυσκολότερη την ταξινόμηση μέσα σε 345 διαφορετικές κλάσεις.

Το τελευταίο σετ δεδομένων που θα εκπαιδεύτει στο τρίτο πείραμα αποτελείται από φωτογραφίες με διαφορετικά είδη λουλουδιών. Θα ακολουθηθεί η κατανομή φωτογραφιών και η διαδικασία που ακολούθησε η (47). Τα δεδομένα αποτελούνται από 102 διαφορε-

τικά είδη λουλουδιών. Το κάθε είδος λουλουδιού συνεπώς και η κάθε κατηγορία που θα πρέπει να το ταξινομήσει το μοντέλο έχει από 40 εώς 258 εικόνες. Οι κατηγορίες για το σετ εκπαίδευσης, για το σετ επικύρωσης και για το σετ ελέγχου είναι 71, 16 και 15 αντίστοιχα. Τα τρία σετ δεδομένων που επιλέγονται για να εκπαιδευτούν και στα τρία διαφορετικά μοντέλα είναι τελείως διαφορετικά.

Αυτό έχει σαν αποτέλεσμα τα μοντέλα να μπορούν να συγκρατήσουν πολύ διαφορετική πληροφορία στο επίπεδο της μετα-μάθησης η οποία θα δείξει πόσο χρήσιμη θα είναι, στην περίπτωση που δεν θα μπορεί να αλλάξει ανάλογα με την κλάση και το πρόβλημα η κάθε παράμετρος του μοντέλου, στις πρώτες δύο περιπτώσεις (μετα-μάθηση σε διαφορετικά πεδία, μετα-μάθηση με μεταβαλλόμενο ποσοστό εκμάθησης) άλλα και στην περίπτωση της μπεϋζιανής προσαρμοσμένης μετα-μάθησης που υπάρχει η δυνατότητα της προσαρμογής των παραμέτρων των μοντέλων ανάλογα με τα χαρακτηριστικά του κάθε προβλήματος και της κάθε κλάσης.

Το πρώτο σύνολο δεδομένων που θα χρησιμοποιηθεί μόνο για το σετ ελέγχου του τρίτου πειράματος λέγεται Traffic Signs και αφορά πινακίδες σήμανσης. Το σύνολο δεδομένων χρησιμοποιείται με την ίδια μορφή με (48). Στο σύνολο του αποτελείται από 43 διαφορετικά είδη πινακίδων και η κάθε διαφορετική κατηγορία αποτελείται από 900 παραδείγματα. Το τελευταίο σετ δεδομένων που αφορά το τρίτο πείραμα αφορά διαφορετικές κατηγορίες ρούχων και ονομάζεται Fashion-MNIST. Τα δεδομένα θα έχουν ακριβώς την ίδια μορφή όπως στην εργασία (49). Στο σύνολο τους τα δεδομένα περιλαμβάνουν 10 διαφορετικές κατηγορίες και η κάθε κατηγορία έχει από 1000 διαφορετικά παραδείγματα.

5.4 Κατασκευή Πειραμάτων

Η διαδικασία κατασκευής των μοντέλων διέπεται πάντα από κάποιες συγκεκριμένες αρχικές συνθήκες.Τα τρία πειράματα στήθηκαν με παρόμοιες αρχικές συνθήκες και για τα τρία μοντέλα. Η αρχιτεκτονική και των τριών μοντέλων βασίζεται σε 4 μπλόκ Συνελικτικών Νευρωνικών Δικτύων. Το κάθε ένα από αυτά αποτελείται από 32 κανάλια. Σε όλα τα μοντέλα ακολουθείται ομαλοποίηση των τιμών των δεδομέων (batch normalization) σύμφωνα με (15). Επίσης πραγματοποείται πρόωρη διακοπή (early stopping) σε όλα τα μοντέλα και για όλα τα πειράματα της διαδικασίας. Ακόμα, ο αριθμός βημάτων κλίσης για όλα τα μοντέλα είναι 5 κατά την διάρκεια της μετα-μάθησης και 10 στον μετα-έλεγχο των μοντέλων.

Ένα σημαντικό κομμάτι αυτής της διαδικασίας είναι η παρομοίωση του προβλήματος με ρεαλιστικά δεδομένα για να είναι πιο δύσκολη η επιτυχία των μοντέλων που θα δοκιμαστούν. Στην προηγούμενη ενότητα αναφέρθηκαν τα σετ των δεδομένων που θα χρησιμοποιηθούν, καθώς και το μέγεθος και ο αριθμός των κλάσεων που διαθέτει το καθένα. Στα πρώτα δύο προβλήματα υπάρχει μόνο ένα σετ δεδομένων εκπαίδευσης

ενώ στο τρίτο υπάρχουν τρία διαφορετικά σετ εκπαίδευσης. Το ζήτημα είναι πως θα επιτευχθεί η ανισορροπία μεταξύ των κλάσεων και των προβλημάτων της διαδικασίας, από την στιγμή μάλιστα που στα δύο πρώτα προβλήματα όλες οι κλάσεις του σετ εκπαίδευσης περιέχουν τον ίδιο αριθμό δεδομένων. Για την επίτευξη αυτής της ανισορροπίας θα ακολουθηθεί η παρακάτω διαδικασία.

Το πρώτο βήμα για να δημιουργηθεί μια κατανομή προβλημάτων $p(\tau)$ θα είναι να επιλεχθούν τυχαία $C = 5$ κλάσεις από το σύνολο των κλάσεων του σετ εκπαίδευσης. Στην συνέχεια με πιθανότητα $p = 0.5$, ο αριθμός των παραδειγμάτων για την κάθε μια από τις πέντε κλάσεις θα δίνεται από $N_c \sim Unif(1, 50)$ ανεξάρτητα για το ποια κλάση αναφέρεται έτσι ώστε να δημιουργηθεί η ανισορροπία δεδομένων μεταξύ των κλάσεων του προβλήματος. Με την άλλη πιθανότητα $p^c = 0.5$ και πάλι ο αριθμός των παραδειγμάτων για την κάθε κλάση θα δίνεται από $N_c \sim Unif(1, 50)$ αλλά σε αυτή την περίπτωση αφορά όλες τις κλάσεις έτσι ώστε να δημιουγηθεί αυτή την φορά η ανισορροπία των δεδομένων μεταξύ των προβλημάτων της διαδικασίας. Ο αριθμός των δεδομένων ελέγχου για την κάθε κλάση θα είναι 15.

Με αυτή την διαδικασία κάθε ένα από τα πειράματα θα προσπαθεί να προσεγγίζει ένα πρόβλημα ρεαλιστικών συνθηκών. Τα υπό προβλήματα που θα αντιμετωπίζει το μοντέλο θα χωρίζονται ανά 100 επαναλήψεις. Συνεπώς, το μοντέλο θα εκπαιδεύεται στις πρώτες 100 επαναλήψεις στην πρώτη περίπτωση με 5 κλάσεις με διαφορετικό αριθμό δεδομένων. Στην δεύτερη περίπτωση, ο αριθμός των δεδομένων για τις 5 αυτές κλάσεις θα είναι διαφορετικός. Οπότε μεταξύ αυτών των δύο προβλημάτων θα υπάρχει διαφορά στο μέγεθος τους και στο πρώτο πρόβλημα θα υπάρχει διαφορα και μεταξύ των παραδειγμάτων της κάθε κλάσης. Με αυτό τον τρόπο θα επιτυγχάνεται η ανισορροπία δεδομένων μεταξύ των κλάσεων ενός προβλήματος αλλά και μεταξύ των προβλημάτων της διαδικασίας.

Τα τρία πειράματα ακολουθούν κάποιες ακόμα αρχικές συνθήκες. Η πρώτη αφορά το μοντέλο της μετα-μάθησης σε διαφορετικά πεδία και πιο συγκεκριμένα το βήμα κλίσης του μοντέλου. Στο πρώτο πείραμα που αφορά τα δεδομένα για το CIFAR 100 το βήμα κλίσης α θα είναι ίσο με 0.5 και στο δεύτερο πείραμα για ένα υποσύνολο δεδομένων του miniImageNet το βήμα κλίσης θα είναι ίσο με 0.1 . Το βήμα κλίσης εξετάστηκε για τις παρακάτω τιμές $\alpha \in (0.01, 0.05, 0.1, 0.5)$.

Η εκπαίδευση των μοντέλων για το πρώτο πείραμα, εγινε για 50 χιλιάδες επαναλήψεις με το μέγεθος της ομαδοποίησης (meta-batch size) να είναι ίσο με 4 . Ακόμα, το ποσοστό μάθησης (outer learning rate) θα είναι ίσο με 0.001 για όλα τα μοντέλα. Το δεύτερο πείραμα αντίστοιχα, έγινε για 80 χιλιάδες επαναλήψεις με μέγεθος ομαδοποίησης 4 και ποσοστό μάθησης ίσο με 0.0001 για όλα τα μοντέλα.

Το τρίτο σετ δεδομένων θα είναι ένα υποσύνολο των δεδομένων που αναφέραμε νωρίτερα για το τρίτο πείραμα. Αρχικά έγινε αλλαγή στο μέγεθος των εικόνων για να είναι όλες

32*32 pixels. Σε αυτό το πείραμα, το μοντέλο της μετα-μάθησης σε διαφορετικά πεδία θα έχει βήμα κλίσης α ίσο με 0.5. Ακόμα, θα τρέξει για 60 χιλιάδες επαναλήψεις με μέγεθος ομαδοποίησης 3 και ποσοστό μάθησης ίσο με 0.001 για όλα τα μοντέλα.

5.5 Αποτελέσματα

Τα αποτελέσματα των τριών πειραμάτων απεικονίζονται παρακάτω σε τέσσερις διαφορετικούς πίνακες. Αρχικά στον πρώτο πίνακα εμφανίζονται τα αποτελέσματα των δύο πρώτων πειραμάτων και στα τρία μοντέλα που χρησιμοποιήθηκαν. Αυτός ο πίνακας αφορά τρεις επαναλήψεις της διαδικασίας ελέγχου όπου η κάθε διαδικασία αποτελείται από 3000 επεισόδια. Στους πίνακες τα αποτελέσματα που παραθέτονται είναι η μέση τιμή αυτών των τριών επαναλήψεων με 0.95 διαστήματα εμπιστοσύνης

Στην πρώτη γραμμή βρίσκεται το μοντέλο της μετα-μάθησης σε διαφορετικά πεδία, στην δεύτερη γραμμή το μοντέλο μετα-μάθησης σε διαφορετικά πεδία με μεταβαλλόμενο ποσοστό μάθησης και στην τρίτη γραμμή το μοντέλο της Μπευζιανής προσαρμοσμένης μετα-μάθησης. Οι δύο πρώτες στήλες του πίνακα αναφέρονται στα αποτελέσματα των δύο σετ δεδομένων του πρώτου πειράματος. Το πρώτο σετ CIFAR 100 που συμμετείχε ,στην διαδικασία εκπαίδευσης και ελέγχου, και το δεύτερο σετ δεδομένων SVHN που συμμετείχε μόνο στην διαδικασία ελέγχου του πρώτου πειράματος. Αντίστοιχα, στην τρίτη και τέταρτη στήλη βρίσκονται τα αποτελέσματα του miniImageNet και του CUB που αφορούν το δεύτερο πείραμα της διαδικασίας. Το πρώτο βρίσκεται στο σετ εκπαίδευσης και στο σετ ελέγχου ενώ το δεύτερο μόνο στο σετ ελέγχου.

Το βασικό συμπέρασμα που απεικονίζεται στον πρώτο πίνακα είναι οτι η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση πετυχαίνει καλύτερα αποτελέσματα ταξινόμησης σε όλα τα σετ δεδομένων. Μάλιστα, στα σετ δεδομένων ελέγχου φαίνεται τα αποτελέσματα της να έχουν μεγαλύτερη διαφορά από τα άλλα δύο μοντέλα. Αυτή η διαφορά είναι λογική από την στιγμή που τα μοντέλα στα σετ δεδομένων που χρησιμοποιούνται μόνο για εκπαίδευση έρχονται αντιμέτωπα με κλάσεις δεδομένων που δεν έχουν εκπαιδευτεί οπότε είναι πολύ δύσκολο να φτάσουν στα επίπεδα επιτυχίας της Μπεϋζιανής προσαρμοσμένης μετα-μάθησης.

Αυτή η διαφορά οφείλεται κυρίως στην ύπαρξη των τριών αυτών τυχαίων μεταβλητών που προσαρμόζουν το μοντέλο ανάλογα με την κλάση και το πρόβλημα που ανήκουν τα δεδομένα. Ακόμα, είναι εύκολο να παρατηρηθεί οτι η αλλαγή του ποσοστού μάθησης στο δεύτερο μοντέλο δηλαδή, η δυνατότητα που δίνει για μεταβαλλόμενο ποσοστό μάθησης δίνει καλύτερα αποτελέσματα σε όλα τα προβλήματα. Το συμπέρασμα που μπορεί να αναχθεί από αυτό τον πίνακα είναι, οτι με την προσαρμογή του σετ δεδομένων σε ένα ρεαλιστικό πρόβλημα τα αποτελέσματα της Μπεϋζιανής προσαρμοσμένης μετα-μάθησης δίνουν μια αύξηση στο ποσοστό ταξινόμησης περίπου στο 5%.

Το συμπέρασμα που προκύπτει ακόμα είναι οτι το σετ ελέγχου του πρώτου πειράματος έχει τα χειρότερα αποτελέσματα από όλα τα σετ δεδομένων. Το γεγονός αυτό μπορεί να οφείλεται στην διαφορά που έχει με το σετ εκπαίδευσης. Ακόμα στο δεύτερο πείραμα το σετ ελέγχου που χρησιμοποιείται είναι ένα σετ που επιλέγεται πολύ συχνα για έλεγχο μοντέλων που έχουν σαν σετ εκπαίδευσης ένα υποσύνολο του miniImagenet.

	CIFAR-FS	SVHN	miniImageNet	CUB
MAML	71.55±0.23	45.17±0.22	66.64±0.22	65.77±0.24
Meta-SGD	72.71±0.21	46.45±0.24	69.95±0.20	65.94±0.22
Bayesian-TAML	**75.15±0.20**	**51.87±0.23**	**71.46±0.19**	**71.71±0.21**

Σχήμα 15: Αποτελέσματα Πρώτου και Δεύτερου Πειράματος μετά από τρία τρεξίματα από 3000 επεισόδια ελέγχου

Ο δεύτερος πίνακας απεικονίζει τα αποτελέσματα του τρίτου πειράματος για τις 3 επαναλήψεις των 3000 επισοδίων. Οι τρείς γραμμές απεικονίζουν ακριβώς όπως και στον προηγούμενο πίνακα τα τρία μοντέλα. Οι τρείς πρώτες στήλες του πίνακα απεικονίζουν τα αποτελέσματα των σετ δεδομένων που χρησιμοποιούνται και στο σετ εκπαίδευσης ενώ οι δύο τελευταίες τα αποτελέσματα των σετ δεδομένων που βρίσκονται μόνο στο σέτ ελέγχου.

Τα αποτελέσματα του τρίτου πειράματος δείχνουν οτι η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση δίνει πολύ καλύτερα αποτελέσματα από τα άλλα δύο μοντέλα. Σε κάποια από τα σετ δεδομένων η διαφορά στα ποσοστά επιτυχίας φτάνουν στο 8%. Σε αυτό το πρόβλημα τα μοντέλα έρχονται πάλι αντιμέτωπα με την ανισορροπία των κλάσεων και την ανισορροπία των υποπροβλημάτων της διαδικασίας. Το γεγονός όμως οτι αυτή τη φορά εκπαιδευέται το μοντέλο σε ακόμα περισσότερες κλάσεις, με μεγαλύτερες διαφορές αφού προέρχονται από τρία διαφορετικά σετ δεδομένων, οδηγεί στο συμπέρασμα οτι είναι λογική η διαφορά της επιτυχίας της ταξινόμησης μεταξύ των τριών μοντέλων.

Γενικότερα από τα αποτελέσματα όλων των μοντέλων σε όλα τα σετ δεδομένων μπορεί να εξαχθεί το συμπέρασμα οτι είναι πιο δύσκολη η ταξινόμηση μεταξύ των συκεκριμένων κλάσεων των μοντέλων και ειδικότερα στην περίπτωση του σετ δεδομένων που αποτελείται από 108 διαφορετικούς τύπους αεροσκαφών. Μια ακόμα παρατήρηση που μπορεί να γίνει είναι ότι σε αυτή την περίπτωση το μεταβαλλόμενο ποσοστό μάθησης που υπάρχει στο μοντέλο που βρίσκεται στην δεύτερη γραμμή του πίνακα

δεν αλλάζει πολύ τα αποτελέσματα στα σετ δεδομένων που χρησιμοποιήθηκαν μόνο για τον έλεγχο των δεδομένων.

	Aircarft	Quickdraw	VGG-Flower	Traffic Signs	Fashion-MNIST
MAML	48.60±0.17	69.02±0.18	60.38±0.16	51.96±0.22	63.10±0.15
Meta-SGD	49.71±0.17	70.26±0.16	59.41±0.27	52.07±0.35	62.71±0.25
Bayesian-TAML	**54.43±0.16**	**72.03±0.16**	**67.72±0.16**	**64.81±0.21**	**68.94±0.13**

Σχήμα 16: Αποτελέσματα Τρίτου Πειράματος μετά από τρία τρεξίματα από 3000 επεισόδια ελέγχου

Οι δύο επόμενοι πίνακες αφορούν την επανάληψη των τριών προηγούμενων πειραμάτων αυτή την φορά με μόνο 1000 επεισόδια. Στους πίνακες τα αποτελέσματα που παραθέτονται είναι τα αποτελέσματα του κάθε μοντέλου με 0.95 διαστήματα εμπιστοσύνης.

Ο πρώτος πίνακας αφορά τα δύο πρώτα πειράματα και ο δεύτερος πίνακας το τελευταίο πείραμα. Τα συμπεράσματα που μπορούν να βγούν είναι παρόμοια με τους παραπάνω δύο πίνακες γεγονός που δέιχνει οτι ακόμα και σε πολύ λιγότερα επεισόδια ελέγχου η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση έχει πολύ καλύτερα ποσοστά ταξινόμησης των κλάσεων σε σχέση με τα δύο άλλα μοντέλα. Αυτή η διαφορά οφείλεται κυρίως στις τυχαίες μεταβλητές που έχουν εισαχθεί στο μοντέλο και έχουν παίξει ρόλο στην προσαρμογή των παραμέτρων ανάλογα με την κλάση και το πρόβλημα της διαδικασίας, αλλά και όπως θα δούμε στα Συμπεράσματα και στον τρόπο κωδικοποίησης των δεδομένων.

	CIFAR-FS	SVHN	miniImageNet	CUB
MAML	72.23±0.67	47.19±0.63	66.95±0.71	66.82±0.73
Meta-SGD	72.93±0.66	47.63±0.73	68.04±0.67	66.45±0.63
Bayesian-TAML	**74.97±0.62**	**52.25±0.68**	**71.27±0.59**	**72.89±0.62**

Σχήμα 17: Αποτελέσματα Πρώτου και Δεύτερου Πειράματος μετά από 1000 επεισόδια ελέγχου

	Aircarft	Quickdraw	VGG-Flower	Traffic Signs	Fashion-MNIST
MAML	48.17±0.57	68.57±0.60	60.68±0.53	52.37±0.78	62.57±0.48
Meta-SGD	51.76±0.61	70.05±0.54	64.28±0.60	50.89±0.1.02	62.83±0.64
Bayesian-TAML	**55.70±0.53**	**72.40±0.50**	**68.39±0.50**	**64.17±0.74**	**67.60±0.47**

Σχήμα 18: Αποτελέσματα Τρίτου Πειράματος μετά από 1000 επεισόδια ελέγχου

5.6 Συμπεράσματα

Στην συγκερκιμένη ενότητα μετά τα αποτελέσματα που αποτυπώθηκαν στους πίνακες ταξινόμησης, θα γίνει εξάγωγη συμπερασμάτων για τα παραπάνω αποτελέσματα. Αρχικά η Μπεϋζιανή προσαρμοσμένη μετα-μάθηση φαίνεται να τα πηγαίνει καλύτερα από τα άλλα δύο μοντέλα και στα εννέα σετ δεδομένων. Αυτό το αποτέλεσμα δεν αλλάζει ούτε όταν υπάρχουν λίγοτερα δεδομένα ελέγχου. Κάτα την διάρκεια της διαδικασίας για να αποτυπωθεί το πρόβλημα σε ρεαλιστικές συνθήκες χρησιμοποιήθηκε αυτή η μέθοδος με τον αριθμό των δεδομένων για την κάθε κλάση στην όποια εκπαιδεύεται το μοντέλο. Όπως γίνεται αντιληπτό στον παρακάτω πίνακα, επιλέγονται πέντε κλάσεις εκ των οποίων τα δεδομένα της κάθε μιας με πιθανότητα 50% επιλέγονται από μια ομοιόμορφη κατανομή από τιμές 1 εώς 50 και με την υπόλοιπη πιθανότητα και οι πέντες κλάσεις έχουν την ίδια τιμή δεδομένων, όπου πάλι αυτή η τιμή είναι μια τιμή από την ομοίομορφη κατανομή από 1 εώς 50.

Ο πίνακας δείχνει τις τιμές της παραμέτρου ω για τις διάφορες τιμές των κλάσεων από την ομοιόμορφη κατανομή. Για μεγαλύτερες τιμές δεδομένων από την ομοιόμορφη κατανομή το ω παίρνει και μικρότερες τιμές. Ακόμα για ίδιο αριθμό δεδομένων μεταξύ των κλάσεων το ω παίρνει ίδιες τιμές. Αυτές οι αλλαγές στις τιμές του ω εμφανίζονται μόνο στο τρίτο μοντέλο που υπάρχουν αυτές οι τυχαίες μεταβλητές που επηρεάζουν την τιμή των παραμέτρων του μοντέλου. Συνεπώς η επιρροή που ασκούν στο μοντέλο είναι σίγουρα θετική και φαίνεται να χρειάζεται η προσθήκη τους.

Μιά ακόμα παρατήρηση που μπορούμε να κάνουμε είναι οτι η τιμή της ω εξαρτάται όχι μόνο από τον αριθμό των δεδομένων κάθε κλάσης, αλλά από τον αριθμό τους σε σχέση μέ των αριθμό όλων των δεδομένων που εμφανίζονται στο κάθε πρόβλημα της διαδικασίας. Αυτό φαίνεται στον πίνακα, όπου για την τρίτη κλάση υπάρχει ο ίδιος αριθμός δεδομένων στο πρόβλημα 9 και στο πρόβλημα 5, αλλά η τιμή του ω είναι διαφορετικη για την τρίτη κλάση στις δύο αυτές περιπτώσεις. Συμπερασματικά λοιπόν η τιμή του ω εξαρτάται από τα δεδομένα της κλάσης δεδομένου του αριθμού των δεδομένων του προβλήματος.

Ο δεύτερος πίνακας αφορά την τυχαία μεταβλητή γ του προβλήματος. Η τυχαία

```
*** Omega for class imbalance ***
         C1  C2  C3  C4  C5        C1    C2    C3    C4    C5
task 1:   1   4   6  17  29 --> 0.444 0.273 0.198 0.056 0.029
task 6:  15  15  15  15  15 --> 0.200 0.200 0.200 0.200 0.200
task 3:   1   4  17  31  35 --> 0.539 0.332 0.068 0.033 0.028
task 7:  13  13  18  20  48 --> 0.292 0.292 0.195 0.166 0.055
task 8:  23  23  23  23  23 --> 0.200 0.200 0.200 0.200 0.200
task 9:  14  20  26  38  43 --> 0.384 0.236 0.174 0.113 0.094
task 5:  14  22  26  33  47 --> 0.394 0.206 0.178 0.138 0.084
task 0:  13  13  28  47  48 --> 0.361 0.361 0.140 0.071 0.068
task 2:  37  37  37  37  37 --> 0.200 0.200 0.200 0.200 0.200
task 4:  49  49  49  49  49 --> 0.200 0.200 0.200 0.200 0.200
```

Σχήμα 19: Οι τιμές που παίρνει η παράμετρος ω για την ανισορροπία των προβλημάτων

μεταβλητή γ προσαρμόζεται σύμφωνα με την θεωρία με το μέγεθός των δεδομένων που υπάρχουν σε κάθε πρόβλημα και είναι μεγαλύτερη όσο μικρότερος είναι ο αριθμός των δεδομένων. Αυτή η διαφορά έγκειται στο ότι προσπαθεί σε προβλήματα που ο αριθμός των δεδομένων είναι μικρότερος να διαδραματίσει μεγαλύτερο ρόλο στην εκμάθηση των δεδομένων η μετα-μάθηση και όχι εκμάθηση των δεδομένων από το συγκεκριμένο πρόβλημα αφού έχει μικρότερο αριθμό δεδομένων από άλλα προβλήματα. Συνεπώς, η πληροφορία που έχει ένα πρόβλημα με λιγότερα δεδομένα είναι μικρότερη.

```
*** Gamma for task imbalance ***
              conv1 conv2 conv3 conv4 dense
task 1: N= 57 0.772 0.775 0.523 0.627 9.438
task 6: N= 75 0.807 0.917 0.834 0.851 4.897
task 3: N= 88 0.785 0.797 0.562 0.658 8.516
task 7: N=112 0.815 0.932 0.895 0.882 4.591
task 8: N=115 0.829 1.001 1.120 1.010 3.469
task 9: N=141 0.831 0.988 1.091 0.990 3.654
task 5: N=142 0.831 0.992 1.104 0.997 3.602
task 0: N=149 0.827 0.961 0.999 0.939 4.094
task 2: N=185 0.853 1.071 1.435 1.162 2.672
task 4: N=245 0.853 1.073 1.443 1.166 2.656
```

Σχήμα 20: Οι τιμές που παίρνει η παράμετρος γ για την ανισορροπία των προβλημάτων

Ο τελευταίος πίνακας αφορά την διαδικασία κωδικοποίησης που ακολουθήθηκε. Στο συγκεκριμένο πίνακα απότυπώνεται η σημασία της δημιουργίας της στατιστικής μελέτης (Statistic Pooling). Αρχικά δίνονται τα αποτελέσματα της Μπεϋζιανής προσαρμοσμένης μετα-μάθησης με χρήση μέσου όρου (Mean Pooling) κατά την κωδικοποίηση. Στην δεύτερη γραμμή φαίνεται η χρήση του μέσου όρου και του αριθμού των δεδομένων και στην τρίτη γραμμή η διαδικασία που χρησιμοποιήθηκε.

Με τον όρο ιεραρχική κωδικοποίηση στον πίνακα αναφέρεται η κωδικοποίηση αρχικά στο κάθε πρόβλημα ξεχωριστά, δηλαδή στον αριθμό των κλάσεων του προβλήματος και στον αριθμό των δεδομένων της κάθε κλάσης του συγκεκριμένου προβλήματος, και στην συνέχεια στην κωδικοποίηση για όλα τα προβλήματα της διαδικασίας μαζί ώστε να γίνεται διαφορετική κωδικοποίηση ανάλογα με το πρόβλημα. Το συμπέρασμα που προκύπτει από τον παρακάτω πείραμα είναι οτι η χρήση της Στατιστικής Μελέτης έιναι πολύ σημαντική για την διαδικασία αλλά ίσως ακόμα σημαντικότερη είναι η κωδικοποίηση πρώτα ώς προς το κάθε πρόβλημα ξεχωριστά και στην συνέχεια ως πρός τα προβλήματα μεταξύ τους.

Meta-training / Meta-test CIFAR-FS / CIFAR-FS	Hierarchical encoding ×	 √
Mean	$73.84_{\pm 0.21}$	$73.69_{\pm 0.21}$
Mean + N	$73.17_{\pm 0.21}$	$74.88_{\pm 0.20}$
Mean + Var. + N	$73.93_{\pm 0.21}$	$\mathbf{75.15}_{\pm \mathbf{0.20}}$

Σχήμα 21: Τα αποτελέσματα της Μπεϋζιανής προσαρμοσμένης μετα-μάθησης ανάλογα με την κωδικοποίηση

References

[1] Open University Course Team *Bayesian statistics*. TheOpenUniversity, Walton-Hall, MiltonKeynesMK76AA, 2007.

[2] Λουκία Μελιγκοτσίδου. *Μπεϋζιανή Συμπερασματολογία*. Σχολή Θετικών Επιστημών Τμήμα Μαθηματικών 2019.

[3] José M. Bernardo. *BAYESIAN STATISTICS*. Departamento de Estadística, Facultad de Matemáticas, 46100–Burjassot, Valencia, Spain, 2003.

[4] Πέτρος Δελλαπόρτας – Παναγιώτης Τσιαμυρτζής. *Στατιστική κατά Bayes*. ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ5.

[5] Brendon J. Brewer. *Introduction to Bayesian Statistics*. `http://creativecommons.org/licenses/by-sa/3.0/deed.en GB.`

[6] Κοκολάκης Γεώργιος , Σπηλιώτης Ιωάννης. *Εισαγωγή στην Θεωρία Πιθανοτήτων και Εφαρμογές*, 2000.

[7] Yaquing Wang, Quanming Yao,James T. Kwok, and Lionel M. Ni. *Generalizing from a Few Examples: A Survey on Few-Shot Learning*. ACM Comput. Surv. 1, 1, Article 1 (March 2020), 34 pages. https://doi.org/10. 1145/3386252

[8] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. *Siamese Neural Networks for One-shot Image Recognition*. Department of Computer Science, University of Toronto. Toronto, Ontario, Canada, 2015.

[9] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. *Matching Networks for One Shot Learning*. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

[10] Jake Snell,Kevin Swersky and Richard S. Zemel. *Prototypical Networks for Few-shot Learning*. `https://arxiv.org/abs/1703.05175v2`

[11] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle and Richard S. Zemel. *META-LEARNING FOR SEMI-SUPERVISED FEW-SHOT CLASSIFICATION*. Published as a conference paper at ICLR 2018

[12] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr and Timothy M. Hospedales. *Learning to Compare: Relation Network for Few-Shot Learning*. `https://arxiv.org/abs/1711.06025`

[13] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao and Jiebo Luo. *Distribution Consistency Based Covariance Metric Networks for Few-Shot Learning*. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2019.

[14] Ricardo Vilalta and Youssef Drissi. *A Perspective View and Survey of Meta-Learning* . Hawthorne, NY., 10532 U.S.A., 2002.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[16] Chelsea Finn, Kelvin Xu, and Sergey Levine. *Probabilistic model-agnostic meta-learning*. In NeurIPS, 2018.

[17] Ian J. Goodfellow, Jonathon Shlens , Christian Szegedy. *EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES*. Published as a conference paper at ICLR 2015.

[18] Adam Santoro, Sergey Bartunov, Matthew Botvinick, DaanWierstra and Timothy Lillicrap. *Meta-Learning with Memory-Augmented Neural Networks*. Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016.

[19] Spyros Gidaris and Nikos Komodakis. *Dynamic Few-Shot Visual Learning without Forgetting*. This work was supported by the ANR SEMAPOLIS project, an INTEL gift, and hardware donation by NVIDIA, 2018.

[20] Qianru Sun, Yaoyao Liu, Tat-Seng Chua and Bernt Schiele. *Meta-Transfer Learning for Few-Shot Learning*.
`https://arxiv.org/abs/1812.02391`

[21] Jake Snell and Richard Zemel. *Bayesian Few-Shot Classification with One-vs-Each Pólya-Gamma Augmented Gaussian Processes*.
`https://arxiv.org/abs/2007.10417`

[22] Tom Silver, Kelsey R. Allen, Alex K. Lew, Leslie Kaelbling and Josh Tenenbaum. *Few-Shot Bayesian Imitation Learning with Logical Program Policies*. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2020.

[23] Eleni Triantafillou, Richard Zemel and Raquel Urtasun. *Few-Shot Learning Through an Information Retrieval Lens*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[24] Andrew L. Maas and Charles Kemp. *One-Shot Learning with Bayesian Networks*. 2009.

[25] Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. *LEARNING TO BALANCE: BAYESIAN META-LEARNING FOR IMBALANCED AND OUT-OF-DISTRIBUTION TASKS.* Published as a conference paper at ICLR 2020.

[26] Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu and Xiaokang Yang. *Variational Few-Shot Learning.* Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October, 2019.

[27] Sachin Ravi and Alex Beatson. *Amortized bayesian meta-learning.* In ICLR, 2019.

[28] Diederik P. Kingma and Max Welling *Auto encoding variational bayes..* In ICLR,
2014.

[29] Jürgen Schmidhuber. *Evolutionary Principles in Self-Referential Learning. On Learning now to Learn: The Meta-Meta-Meta...-Hook.* PhD thesis, Technische Universität München,
1987.

[30] Sebastian Thrun and Lorien Pratt. *Learning to Learn.* Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-8047-9,
1998.

[31] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. *Task dependent adaptive metric for improved few-shot learning.* In NeurIPS, 2018.

[32] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. *Few-shot image recognition by predicting parameters from activations.* In CVPR, 2018.

[33] Sachin Ravi and Alex Beatson. *Amortized bayesian meta-learning.* In ICLR, 2019.

[34] Sachin Ravi and Hugo Larochelle. *Optimization as a model for few-shot learning.* In ICLR, 2017.

[35] Yoonho Lee and Seungjin Choi. *Gradient-based meta-learning with learned layerwise metric and subspace.* In ICML, 2018.

[36] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. *Deep sets.* n NIPS, 2017.

[37] Harrison Edwards and Amos Storkey. *Towards a neural statistician.* In ICLR, 2017.

[38] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. *Metalearning probabilistic inference for prediction.* In ICLR, 2019.

[39] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. *Meta-learning with latent embedding optimization.* In ICLR, 2019.

[40] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. *Fast and flexible multi-task classification using conditional neural adaptive processes.* NeurIPS, 2019.

[41] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. *Meta-sgd: Learning to learn quickly for few shot learning..* arXiv preprint arXiv:1707.09835, 2017.

[42] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. *Meta-learning with differentiable closed-form solvers.* In ICLR, 2019.

[43] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. *Reading digits in natural images with unsupervised feature learning.* 2011.

[44] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. *Reading Fine-grained visual classification of aircraft.* arXiv preprint arXiv:1306.5151, 2013.

[45] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. *Meta-dataset: A dataset of datasets for learning to learn from few examples.* In ICLR, 2020.

[46] Takashi Kawashima Jongmin Kim Jonas Jongejan, Henry Rowley and Nick Fox-Gieg. *The quick, draw! – a.i. experiment.* 2016. URL http://quickdraw.withgoogle.com.

[47] Maria-Elena Nilsback and Andrew Zisserman. *Automated flower classification over a large number of classes.* In 2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing, pp. 722–729. IEEE, 2008.

[48] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. *Detection of traffic signs in real-world images: The german traffic sign detection benchmark.* In The 2013 international joint conference on neural networks (IJCNN), pp. 1–8. IEEE, 2013.

[49] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*. arXiv preprint arXiv:1708.07747, 2017

[50] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. *Bayesian model-agnostic meta-learning*. Advances in Neural Information Processing Systems, pages 7332–7342, 2018.