Η εκτεταμένη χρήση των υπολογιστικών συστημάτων και των δικτύων στις μέρες μας έχει οδηγήσει στη διάχυση της πληροφορίας σε πολλές επιμέρους αυτόνομες συλλογές δεδομένων. Συνεπώς αποκτά ιδιαίτερο ενδιαφέρον η συσχέτιση των συλλογών δεδομένων αυτών με σκοπό την ανακάλυψη κοινών οντοτήτων που ενυπάρχουν σε αυτές. Η επιστημονική περιοχή που ασχολείται με αυτό το αντικείμενο ονομάζεται Διασύνδεση Εγγραφών (Record Linkage)
.

Στις περιπτώσεις που οι οντότητες που περιέχονται στις συλλογές δεδομένων έχουν κάποιο μοναδικό χαρακτηριστικό (πχ το πρωτεύον κλειδί σε μια εγγραφή βάσης δεδομένων) τότε μπορούν να εφαρμοστούν ντετερμινιστικοί αλγόριθμοι για την σύνδεση εγγραφών (Deterministic Record Linkage). Αρκετά συχνά όμως τα δεδομένα δεν περιγράφονται από κάποιο τέτοιο μοναδικό χαρακτηριστικό, οπότε η συσχέτιση των εγγραφών γίνεται με μη-ντετερμινιστικό τρόπο. Σε αυτή την περίπτωση, σημαντική συνεισφορά προσφέρει η εργασία του Winkler
, ο οποίος παρουσιάζει την
Πιθανοτική Σύνδεση Εγγραφών
(Probablistic Record Linkage). Η σύγκριση των εγγραφών πραγματοποιείται με την χρήση του διανύσματος σύγκρισης {tex}gamma{/tex}, το οποίο ορίζεται αυθαίρετα και έχει ως συνιστώσες κάποια από τα πεδία των εγγραφών που συγκρίνονται. Για ένα δεδομένο διάνυσμα {tex}gamma{/tex}, υπολογίζονται οι a posteriori πιθανότητες {tex}P(gamma|M){/tex}, δηλαδή η πιθανότητα όταν το διάνυσμα {tex}gamma{/tex} ικανοποιείται οι εγγραφές όντως να ταιριάζουν και {tex}P(gamma|U){/tex}, δηλαδή η πιθανότητα όταν το διάνυσμα {tex}gamma{/tex} ικανοποιείται οι εγγραφές να μην ταιριάζουν, στο σύνολο του dataset και από αυτές προκύπτει ο λόγος σύνδεσης {tex}R = frac{P(gamma|M)}{P(gamma|U)}{/tex}.

Στη συνέχεια, μπορεί να κατασκευαστεί ένας απλός κανόνας απόφασης (decision rule), ο οποίος τοποθετεί τις συγκρινόμενες εγγραφές σε 3 κατηγορίες, ανάλογα με την τιμή του λόγου σύνδεσης
:

1. Ταίριασμα: Αν ο λόγος σύνδεσης είναι πάνω από το κατώφλι ταιριάσματος (match thresold)
2. Μη Ταίριασμα: Αν ο λόγος σύνδεσης είναι χαμηλότερος από το κατώφλι

μη-ταιριάσματος (non-match threshold)
3. Πιθανό Ταίριασμα: Αν ο λόγος σύνδεσης βρίσκεται μεταξύ των δύο κατωφλίων

Οι “a posteriori” πιθανότητες {tex}P(gamma|U){/tex} και {tex}P(gamma|U){/tex} καθώς και τα αντίστοιχα κατώφλια απόφασης και μη-απόφασης μπορούν να προσεγγιστούν με μια σειρά από τεχνικές, όπως πχ οι Expectation-Maximization αλγόριθμοι. Μπορούν όμως να ειδωθούν και ως ένα πρόβλημα ταξινόμισης (classification), όπου το ζητούμενο είναι να κατασκευαστεί ένας ταξινομητής (classifier), ο οποίος θα παρέχει μια εκτίμηση των εν λόγω τιμών.

Στην εργασία του Minton , κατασκευάζει έναν ταξινομητής που βασίζεται σε Support Vector Machines
. Οι προς σύγκριση εγγραφές απεικονίζονται στην είσοδο του SVM με την χρήση ενός learned distance metric που προτείνει ο συγγραφέας. Το όλο σύστημα εμφανίζει ενθαρρυντικά αποτελέσματα στα datsets στα οποία δοκιμάζεται.

Σκοπός αυτής της εργασία είναι:

1. Να εξεταστεί η απόδοση της συγκεκριμένης μεθοδολογίας στο δημόαια διαθέσιμο Record Linkage Comparison Dataset

(από το
UCI Repository
) καθώς και σε ένα μη-δημόσια διαθέσιμο dataset που κατέχει το εργαστήριο.

2. Να βρεθούν οι αλλαγές που μπορούν να γίνουν στα συστατικά στοιχεία του συστήματος έτσι ώστε τα αποτελέσματα να βελτιωθούν στον καλύτερο δυνατό βαθμό.