

Tra passato e futuro:

l'Intelligenza Artificiale per conoscere i marchi storici

Gran parte delle attività del progetto Matosto consiste nella riflessione condotta da specialisti, giovani, studenti, sul materiale dell'archivio storico Matosto. Si tratta di riflessioni condotte da esseri umani sulla base delle loro conoscenze ed esperienze sulla materia. In questo breve appunto si introducono, a completamento ed arricchimento, elaborazioni dell'Archivio Matosto condotte con metodologie di calcolo automatico, in linea con l'attuale stato dell'arte nell'ambito del Machine e Deep Learning.

I due termini di Machine Learning (ML) e Deep Learning (DL) sono diventati oggi di uso corrente in riferimento a particolari e specifici classi di algoritmi inquadrati nel contesto della più ampia materia della Intelligenza Artificiale, termine ampio, a volte generico, spesso fonte di accese discussioni sulla relazione con l'intelligenza naturale umana e quindi non sempre amato dagli addetti ai lavori. Il Machine Learning, cioè la capacità delle macchine di apprendere, consiste nella capacità di estrarre dell'informazione da archivi e dati pre-esistenti, e di usare questa informazione per organizzare e classificare in qualche modo l'informazione di partenza in assenza di indicazioni puntuali da parte dell'utilizzatore. Concettualmente i programmi di ML possono essere visti come modi di impiego particolarmente sofisticato degli strumenti classici della statistica. All'interno del Machine Learning si riconosce una ulteriore sottoclasse di algoritmi definita Deep Learning caratterizzato dall'impiego di reti neurali organizzate su più livelli e da una conseguente più potente capacità di apprendimento anche in assenza di qualsiasi organizzazione dei dati in ingresso.

All'interno del progetto Matosto Stefano Frache ha svolto due lezioni magistrali che hanno esaminato come la produzione cinematografica e multimediale degli ultimi decenni abbia letto e interpretato i temi dell'intelligenza artificiale, delle macchine pensanti e della "vita sintetica". La presenza sui media di queste tematiche è ormai così diffusa che l'opinione sulle nuove tecnologie, sui suoi vantaggi attesi e sui rischi possibili. Ciò da un lato ha reso oggetto di pubblico interesse tematiche scientifiche e tecnologiche poco presenti nei normali curricula scolastici e formativi; dall'altro non sempre viene approfondita la differenza profonda che intercorre tra il modo di operare dell'intelligenza umana e delle macchine cosiddette "intelligenti". In generale, la rappresentazione dei media tenda a considerare degni di particolare attenzione e ammirazione risultati ottenuti dalle macchine che sono, invece, relativamente semplici per le stesse, sottovalutando, invece, abilità nostre proprie che, apparentemente più semplici, sono in realtà difficilissime da simulare con il calcolo. Ad esempio, è relativamente facile per una macchina

riconoscere il volto di una persona anche in mezzo a persone simili, mentre è quasi impossibile mantenere in equilibrio un robot mentre scende in un bosco accidentato. Non esistono camerieri robotici ma il software di gestione delle mail usa raffinatissimi algoritmi ed una enorme potenza di calcolo per contenere lo spam.

La base dati Matosto nelle sue componenti testuali e grafiche è un candidato quasi naturale per testare la reale portata di queste tecnologie, individuando limiti e concrete possibilità anche in vista di successivi approfondimenti.

Oggi sono disponibili sul mercato una molteplicità di software e di piattaforme professionali che contengono modelli di apprendimento e di classificazione concepiti per essere applicabili a enormi quantità di dati, i cosiddetti big data.

Sono anche disponibili, e spesso liberamente scaricabili dalla rete, piattaforme open source non meno complesse che svolgono le stesse funzioni e che nascono dalla collaborazione tra Università, società di consulenza, imprese e singoli studiosi normalmente rilasciate in modalità Open Source. Ci sono anche siti che esaminano e comparano le molte decine di software esistenti illustrandone pregi e difetti. Una approfondita trattazione di questa materia esula dall'obiettivo di questa nota, preme solo rilevare come, ormai, il numero di persone che lavorano alla realizzazione di queste piattaforme è un fenomeno che coinvolge milioni di specialisti. Costituisce, quindi, un terreno non solo di interesse di prospettiva ma, ormai, una realtà. Realtà ancora poco presente e sottovalutata nel nostro paese.

Per analizzare i dati di Matosto si è utilizzato una di queste piattaforme: Knime. Questo piattaforma contiene migliaia di moduli di cui possono essere definite le proprietà e le modalità operative e che, collegati tra di loro in modo grafico, trasformano il processo di analisi dei dati in un grafo in cui ciascun nodo svolge una funzione specifica. Qualora si vogliano adottare algoritmi speciali non disponibili nei moduli già predisposti e scaricabili Knime può integrarsi con ambienti di programmazione quali Java, Python, R, ecc.

Nel nostro caso il lavoro svolto utilizza solo una minima parte dei numerosi moduli disponibili con la piattaforma. In particolare sono stati usati: i moduli di lettura e scrittura dei dati in formato Excel e delle immagini in formato Jpeg; i moduli per la manipolazione delle tabelle, dei contenuti testuali e delle immagini; un modulo di Machine Learning per la classificazione dei testi e due moduli di Deep Learning per l'apprendimento e la classificazione delle immagini.

Il tentativo di applicare i moduli di ML e DL ai dati nativi ha evidenziato la presenza di anomalie ed incongruenze dovute a sviste o a banali errori di caricamento sfuggiti in fase di rilevazione dei dati (sono circa 35.000 i documenti presi in esame e sintetizzati in un foglio Excel di altrettante righe). Sono stati quindi utilizzati moduli Knime per rilevare in modo sistematico tali anomalie, correggerle, in parte con moduli Knime ed in parte manualmente e predisporre i dati alla successiva elaborazione dei moduli di ML e DL.

Incidentalmente sottolineiamo come la parte di lavoro necessario per l'organizzazione dei dati e il controllo di correttezza degli stessi sia stimato, nelle pubblicazioni specialistiche, nell'80% del lavoro complessivo. Il restante 20% viene impiegato per l'analisi dei dati e la costruzione di modelli di ML. Nel nostro caso non siamo lontani da questa percentuale.

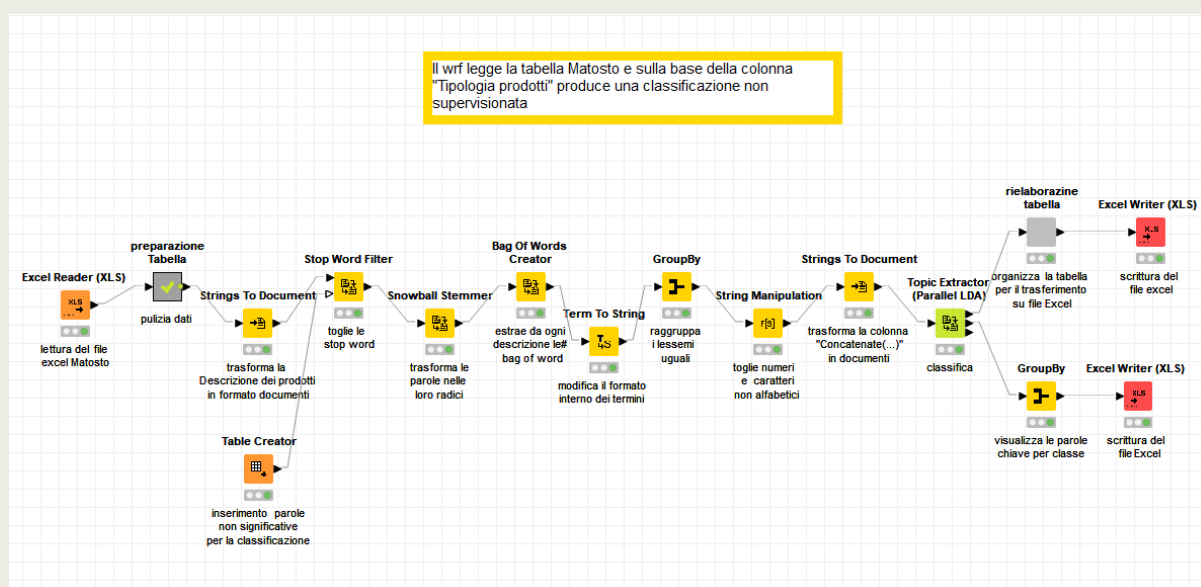
Il modulo di classificazione dei testi è stato utilizzato per produrre una classificazione dei marchi (sulla base dei prodotti associati) più compatta e completa rispetto alla classificazione presente nei dati iniziali. Nei marchi registrati dal 1926 al 1941 non è presente alcuna classificazione. Dopo il 1941 la classificazione fa riferimento alla tabella allegata al Regio Decreto 21 giugno 1942, n. 929 con la suddivisione dei prodotti in 49 classi. Essa è successivamente stata sostituita, in virtù della Legge 10 aprile 1954, n. 129, con altra classificazione identica alla classificazione dei marchi internazionali, che all'epoca comprendeva 34 classi soltanto dedicate ai prodotti. Dal 1 febbraio 1960 si sono susseguite le nuove e sempre più articolate edizioni della Classificazione internazionale di Nizza, definita nel 1957 dall'organizzazione World Intellectual Property Organization (WIPO); ogni cinque anni fino al 2013, quando le nuove edizioni hanno incominciato a susseguirsi annualmente, giungendo oggi a distinguere i prodotti in 34 classi ed i servizi in ulteriori 11 classi. In questo quadro di disomogeneità delle classificazioni utilizzate nel tempo e di eccessivo numero di classi (non meno di 34 fino a 49) ci si è orientati verso l'introduzione di una nuova classificazione organizzata su un numero minore di classi e omogenea per tutto l'insieme dei marchi.

Si è scelto l'utilizzo di una metodologia di apprendimento non supervisionata. Il raggruppamento dei testi in classi avviene cioè senza che fornire alla macchina indicazioni specifici sui criteri di classificazione ma, a partire dalla descrizione delle tipologie di prodotto associate a ciascun marchio il software, basandosi sulla frequenza dei termini e delle loro combinazioni all'interno di ogni singola descrizione, suddivide l'insieme dei marchi in un definito numero di classi stabilito a priori.

In estrema sintesi, il foglio Excel della tabella di lavoro Matosto viene trasformato in una tabella interna a Knime. La tabella è successivamente manipolata per fornire le informazioni ad un modulo di classificazione. Ad ogni descrizione di tipologia di prodotto vengono tolte le cosiddette "stop word", cioè le parole ricorrenti in ogni testo italiano non utili ai fini della classificazione

(articoli, preposizioni, congiunzioni, ecc), e con esse le parole molto ripetute che non sono adatte a generare differenze significative tra le descrizioni (ad esempio il termine “prodotto” che vale per quasi tutte le descrizioni), poi dai termini sono sostituiti dai loro lessemi (la radice significativa) per renderli indipendenti da variazioni quali plurale o singolare, femminile o maschile, ecc. A questo punto la descrizione della tipologia del prodotto si presenta come un elenco dei lessemi delle parole significative. Le descrizioni sono quindi trasferite al classificatore utilizzato, basato sull’algoritmo LDA (Latent Dirichlet Allocation) a cui vengono forniti anche alcuni parametri quali il numero di parole da usare per la classificazione e il numero di classi cui ricondurre le diverse tipologie di prodotto. Si noti che si definisce il numero di parole, non si individuano le parole ritenute più significative: questo è il lavoro che compie il classificatore, che sceglie i termini in modo da rendere massima la distanza nello spazio dei dati tra le diverse classi.

Il workflow elaborativo è qui rappresentato:



Sono state operate classificazioni scegliendo diverse coppie di numeri di parole e di classi. Qui di seguito sono presentati alcuni risultati con la numerosità delle classi distribuite in serie storiche.

Nel caso riportato si è richiesto al modulo classificatore di ripartire l’insieme dei marchi in 5 classi usando 4 parole come discriminanti.

Le parole individuate (o più precisamente i lessemi) che caratterizzano ogni classe sono le seguenti:

Group table - 3:221 - GroupBy (visualizza le parole)

File Edit Hilite Navigation View

Table "default" - Rows: 6 Spec - Columns: 2 Properties Flow Variables

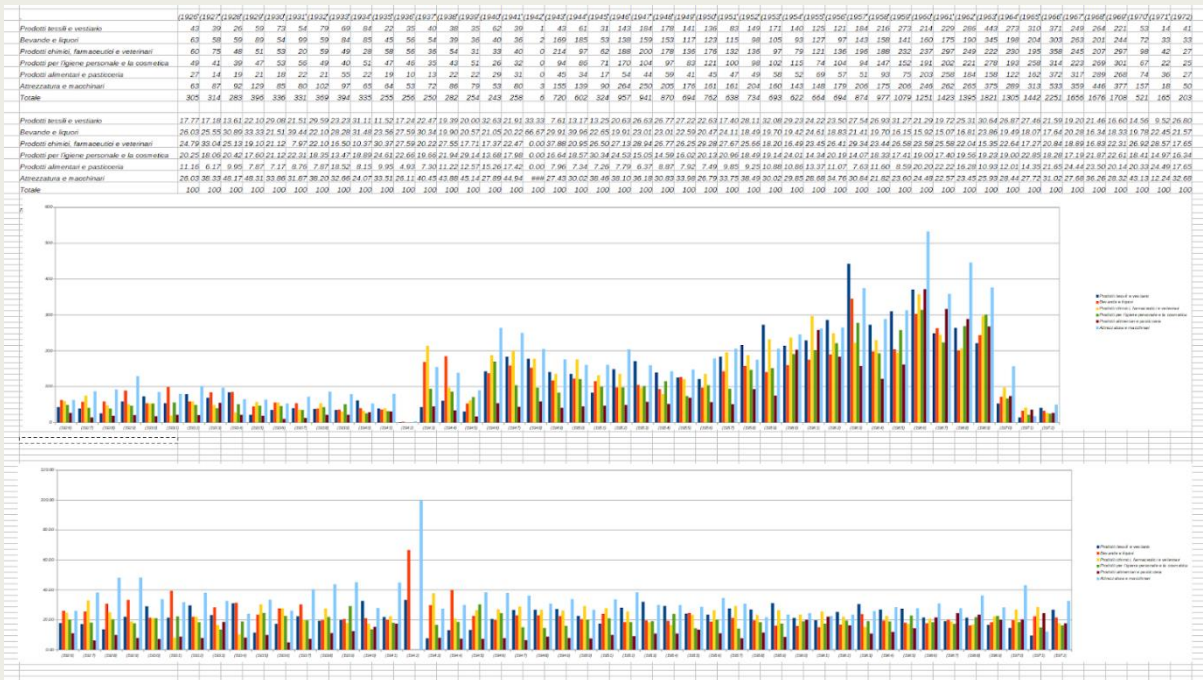
Row ID	Topic id	Concatenate(Term)
Row0	topic_0	tess, fil, scarp, pantofol
Row1	topic_1	liquor, vin, bevand, alcol
Row2	topic_2	farmaceut, medicinal, chimic, veterinar
Row3	topic_3	sapon, macchin, cosmet, profumer
Row4	topic_4	commest, biscott, pasticcer, caff
Row5	topic_5	metall, accessor, macchin, material

tradotte nelle seguenti classi:

topic_0	tess, fil, scarp, pantofol	Prodotti tessili e vestiario
topic_1	liquor, vin, bevand, alcol	Bevande e liquori
topic_2	farmaceut, medicinal, chimic, veterinar	Prodotti chimici, farmaceutici e veterinari
topic_3	sapon, macchin, cosmet, profumer	Prodotti per l'igiene personale e la cosmetica
topic_4	commest, biscott, pasticcer, caff	Prodotti alimentari e pasticceria
topic_5	metall, accessor, macchin, material	Attrezzatura e macchinari

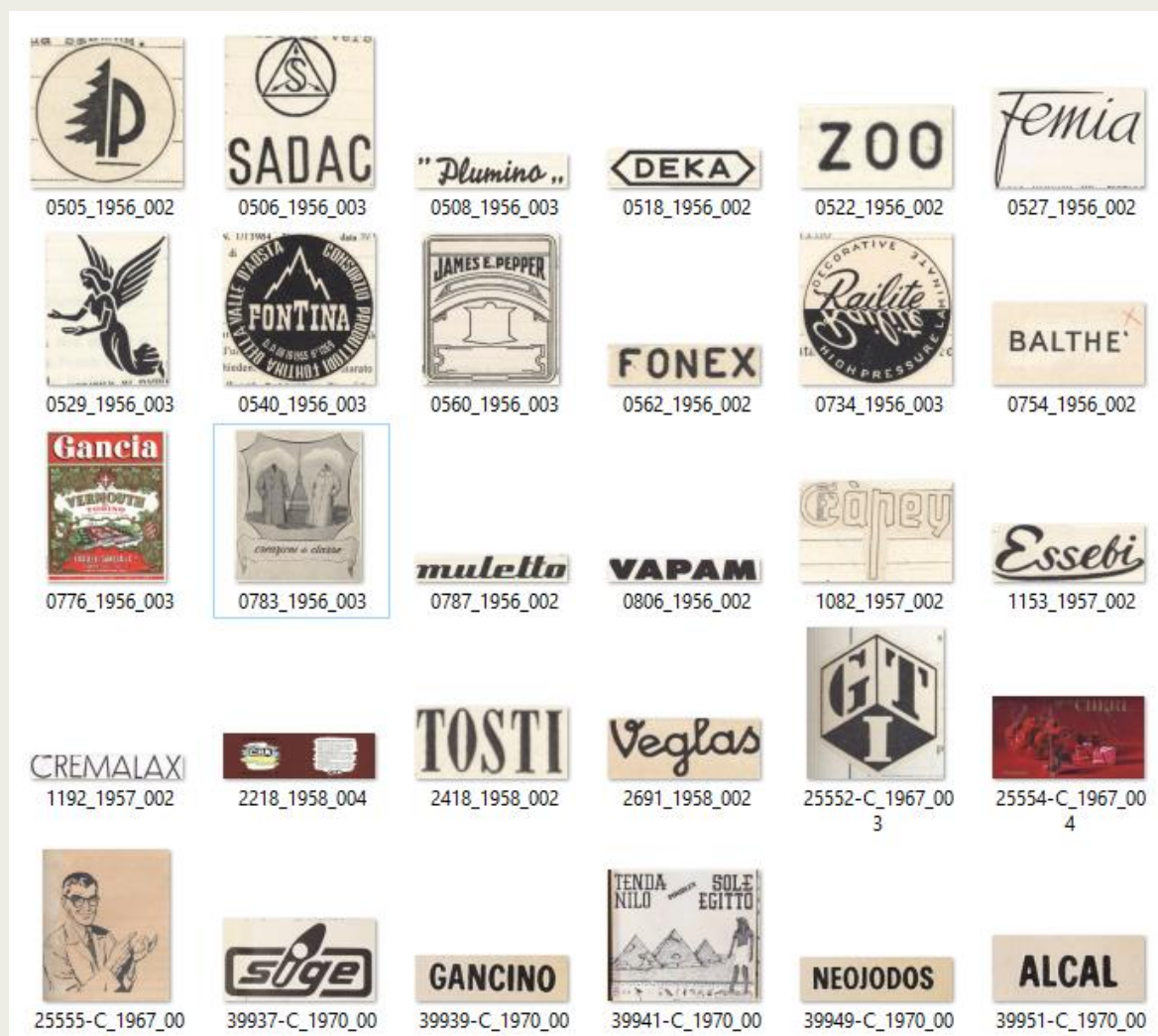
Il risultati ottenuti suddivisi in classi e organizzati per anno sono stati quindi scritti in un foglio Excel di cui riportiamo lo screenshot che rappresenta la tabella dei valori assoluti dei marchi corrispondenti alla coppia classe-anno, poi gli stessi valori in percentuale ed infine gli istogrammi che corrispondono alle due tabelle :

MaToSto®



In generale aumentando il numero di classi e di parole si ottengono classificazioni coerenti con quelle più compatte. Ovviamente in un numero di casi percentualmente modesto ma numericamente non irrilevante la classificazione operata dal sistema è discutibile e comunque diversa da quella effettuata in sede di registrazione dei marchi; registrazione che peraltro associa ai singoli marchi una molteplicità di classificazioni, rendendole di scarsa significatività.

Per la classificazione delle immagini dei marchi le problematiche poste riguardavano la distinzione tra marchi verbali e marchi figurativi. Ogni marchio è definito “Verbale”, cioè costituito da una o più parole, o “Figurativo” caratterizzato cioè da un disegno astratto o meno e/o da una grafica specifica. A tutti i marchi è associata una immagine che riporta la parola piuttosto che il disegno e la grafica specifica. Di seguito alcuni esempi.

MaToSto ®

I marchi sono stati classificati dagli specialisti impegnati nella fase di digitalizzazione dei documenti di registrazione messi a disposizione dalla Camera di commercio. Questa classificazione in molti casi presenta oggettivi elementi di incertezza. Vi sono marchi molto semplici che contengono soltanto una o più parole delle quali il corpo tipografico è per nulla o poco significativo. Vi sono disegni a colori molto ricchi o marchi che contengono simbologie astratte e complesse, forme geometriche, ecc. Vi è, inoltre, un gruppo di marchi nei quali è significativamente presente la componente verbale, accompagnata però da elementi grafici nei corpi tipografici usati, nell'impaginazione, ecc, che rendono non immediatamente evidente la classe di appartenenza.



Ci si è posti quindi l'obiettivo di completare la classificazione e di estrarre i casi di classificazione errata o incerta, attraverso gli strumenti di Deep Learning messi a disposizione dalla piattaforma Knime.

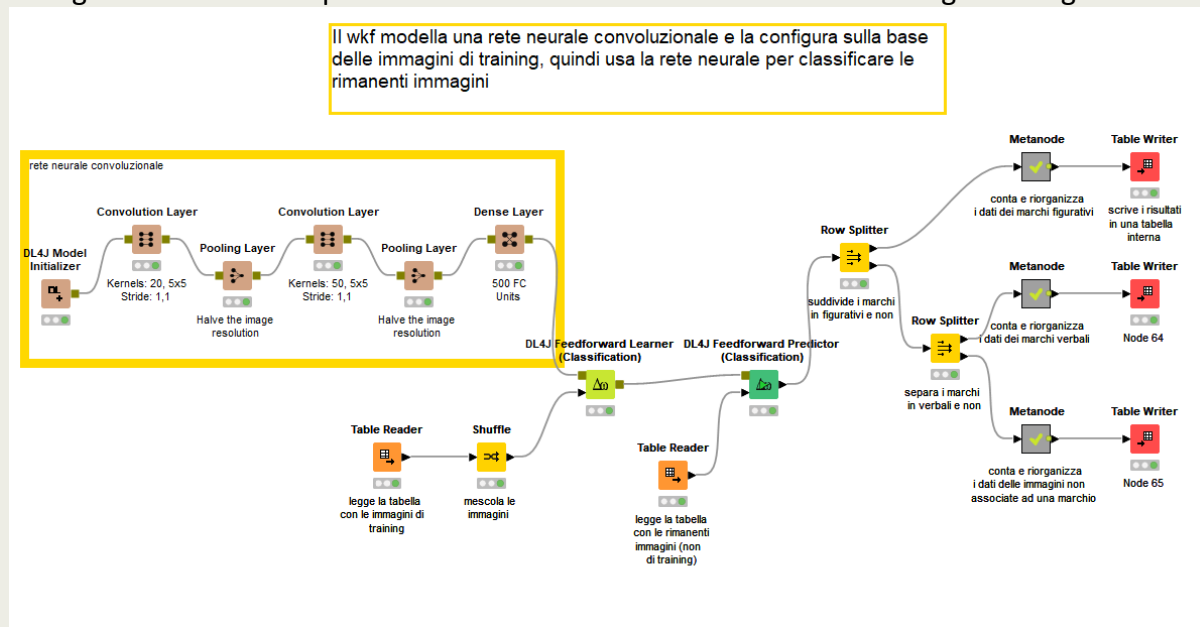
La fase di preparazione dei dati è consistita nell'estrazione, all'interno delle directory contenenti in formato Jpeg le immagini delle domande di registrazione, le sole immagini pertinenti, relative cioè alla immagine del marchio. Dato il numero di file contenute nelle directory (x file organizzati in y directory per z immagini di marchi) questo lavoro è stato possibile grazie all'utilizzo di strumenti automatici, nel nostro caso disponibili all'interno della piattaforma Knime. Non è stato possibile però in questa fase assicurarsi che tutte le immagini estratte fossero esclusivamente riferite ai marchi. Sono rimaste quindi immagini non pertinenti. Le immagini sono state quindi convertite dal formato Jpeg ad un formato interno Knime, normalizzate (stessa dimensione) e quindi ridotte drasticamente nel numero di pixel. La scelta del numero dei pixel da utilizzare è un compromesso tra il carico elaborativo che immagini troppo avrebbero generato rispetto alle risorse di calcolo disponibili e la qualità del risultato che degrada con la riduzione dei pixel. In particolare, si è scelto di operare con una rappresentazione dei marchi di 25 punti in verticale e 100 punti in orizzontale per tener conto del fatto che, sovente, i marchi si presentano in formato rettangolare con più informazione sull'asse x rispetto a quella contenuta sull'asse y.

Il sistema di classificazione delle immagini individuato prevede una fase di apprendimento operata da uno specifico modulo di DL, sulla base di un certo numero di immagini di marchi già correttamente classificate. Il modulo modella e parametrizza opportunamente una rete neurale convoluzionale da usare poi per la classificazione (in questo consiste l'apprendimento). La rete neurale configurata è trasferita poi ad un secondo modulo di DL detto di testing che provvede alla classificazione delle immagini non utilizzate per l'apprendimento. Ciò ha permesso di classificare le immagini senza classificazione nativa e di estrarre i casi di classificazione errata o non pertinente. Il risultato finale è stato ottenuto reiterando alcune volte la sequenza training - testing, utilizzando i casi di classificazione corretta del testing per incrementare il set di immagini dato in pasto alla fase di training. La classificazione ultima è stata realizzata con circa 33.600 immagini, delle quali circa 4400 usate come training. Sulle rimanenti circa 29.000 è stata operata la classificazione. La classificazione delle immagini di training è avvenuta come parte del processo di modellazione, non utilizzando quindi le classificazioni già presenti in archivio.

Le 29.000 immagini risultano distinte in circa 12.000 marchi figurativi e circa 17.000 verbali. Dei circa 12.000 figurativi circa 1.500 differiscono dalla classificazione già presente in Matosto. Dei circa 17.000 verbali circa 1.000 risultavano classificati come figurativi in Matosto. Un esame delle discrepanze è indubbiamente interessante. Nel caso dei 1.500 figurativi difforni dalla classificazione Matosto sembra in molti preferibile la classificazione automatica. Nei 1.000 verbali,

già classificati figurativi in Matosto, la valutazione si inverte. Test effettuati con diversi set di training hanno dato risultati simili.

Di seguito è illustrato la porzione del workflow relativo alla fase di training e testing.



I risultati ottenuti hanno evidenziato l'opportunità di aumentare le classi nelle quali suddividere le immagini, aggiungendo una terza classe "verbale-figurativo" alle precedenti. Si è proceduto quindi in questa direzione fornendo al modulo di training tre gruppi di esempi, verbale, figurativo, verbale-figurativo, sui quali "imparare" i criteri di classificazione, utilizzati poi dal testing. Di questo e di altre analisi per brevità non si riportano in questa sede tabelle analitiche. Va però riportato che i risultati sono decisamente buoni anche se permane un'area di ambiguità ai confini tra le classi; comunque limitata se si pensa sia alle caratteristiche delle immagini trattate (in alcuni casi di incerta classificazione anche da parte di un operatore umano), al limitato numero di immagini utilizzate in fase di training (in letteratura si parla di insiemi di training dalle centinaia di migliaia di casi in su) e alla ridotta risoluzione delle immagini (2500 pixel per immagine). Ad esempio il software classifica come marchi verbali immagini di oggetti allineati, interpretando la serie di oggetti come serie di lettere.

Solo un significativo approfondimento dei modelli utilizzati ed il completamento dell'archivio potranno permettere di introdurre queste algoritmi come parte dell'interfaccia di ricerca sulla base dati.

Giova però rilevare che l'utilizzo delle stesse immagini ridimensionate nel numero di pixel produce in classificatori di immagine professionali e già pre-trained risultati imprevedibili, sovente comici. Ad esempio, sono disponibili in rete modelli scritti in linguaggio Python che tendono a classificare i marchi che contengono immagini rotonde come tombini o medaglioni. L'utilizzo di software pre-trained contenuto in un potentissimo strumento non opensource, cioè Mathematica, disponibile presso ISMEL, produce risultati analogamente curiosi ed inattendibili. Il software pre-trained su un numero notevolissimo di immagini distingue molto bene volti, le diverse specie di animali, ma è pensato per interpretare dati provenienti da macchine fotografiche o telecamere, non materiali d'archivio. Qualora su questo software venisse effettuato un training analogo a quello effettuato su Knime, ci si attende un risultato comparabile. Questo ed altri approfondimenti potrebbero essere oggetto di una ulteriore evoluzione del progetto.

Le prospettive di utilizzo di ML e DL negli archivi delle imprese e della P.A.

Il bisogno di strumenti di analisi dati e costruzione di modelli predittivi nell'attività di ricerca e nella gestione delle imprese è spesso risolto con spreadsheet, primo tra tutti Excel e pacchetti statistici ad hoc. Si tratta in genere di strumenti molto efficaci per affrontare e risolvere problematiche anche complesse basate però sull'utilizzo di informazioni sufficientemente omogenee e strutturate. Il lavoro da noi svolto mostra come in situazioni specifiche, con volumi informativi significativi, informazioni disomogenee (dati, testi, immagini) e non strutturate, sia possibile utilizzare con costi modesti molti altri strumenti disponibili in rete, sovente ritenuti dal nostro mondo imprenditoriale confinati agli ambiti specialistici o accademici.

Gli archivi della P.A. e delle imprese costituiscono in questa luce un giacimento quasi inesauribile di informazioni da utilizzare. Questa prima esperienza condotta sui marchi registrati dalla CCIA di Torino lascia intravedere prospettive promettenti proprio in considerazione delle caratteristiche dei dati elaborati. Da un lato, infatti, si avevano i metadati descrittivi originari, dall'altro un archivio "ibrido" costituito dalla scansione dei documenti (sia grafici, sia iconografici, sia meramente scritti) corredato da nuovi metadati attribuiti dagli archivisti in fase di scansione dei singoli documenti. Questo ha permesso di saggiare le capacità di elaborazione di una macchina "istruita" dai criteri di classificazione adottati dalle persone umane consentendo, al tempo stesso, di verificare la coerenza delle singole schede e, soprattutto, di proporre nuovi raggruppamenti tipologici determinati esclusivamente dalle ricorrenze dei gruppi di lessemi significativi. La significatività del risultato ottenuto (il raggruppamento automatico nelle cinque classi merceologiche) depone a favore di una evidente efficacia dello strumento. Di contro, l'analisi figurativo/verbale dei marchi evidenzia la necessità, quando si lavora su archivi non fotografici, di istruire preliminarmente la macchina per evitare che le forme vengano equivocate sulla base

dell'esperienza di soggettazione delle fotografie contemporanee, esperienza fondamentale, ma non esaustiva dell'universo iconografico disponibile.

La gratuità (o i costi molto contenuti) di strumenti di elaborazione molto potenti unita alla disponibilità nelle imprese e nella P.A. di sterminati archivi di documenti (scritti, iconografici e fotografici) rende realizzabili progetti di "sfruttamento" di tali archivi a scopi sia culturali ed educativi (la memoria di quello che è stato, come è stato veramente) sia per la programmazione delle attività future (il marketing dei prodotti, la progettazione dei nuovi servizi della P.A. e delle infrastrutture...).

Il punto chiave è, come sempre, l'importanza della formazione delle persone e l'apertura verso l'innovazione. Deve formarsi negli anni futuri una nuova leva di archivisti e documentalisti in grado di avere con gli strumenti di ML e DL un rapporto di collaborazione dialettica, evitando gli atteggiamenti di diffidenza preconcepita ("la macchina non può interpretare e discernere criticamente") così come, all'opposto, la delega acritica ("è così perché l'ha detto la macchina"). Davanti a noi si estende una sterminata "nuova frontiera", ma la condizione necessaria alla sua conquista è che sia libera e attraversabile. Il che vuol dire, tradotto nell'ambito degli archivi contemporanei delle imprese e della P.A., che tali archivi vengano aperti (tranne i documenti amministrativi riservati e/o strettamente personali o i segreti industriali) alla consultazione massiva oggi possibile con l'ausilio della potenza di calcolo oggi finalmente accessibile a molti.