UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA

Corso di Laurea in INGEGNERIA GESTIONALE

REGOLE DI ASSOCIAZIONE PER L'ANALISI DI AFFIDABILITÀ DEI SISTEMI

ASSOCIATION RULES FOR SYSTEM RELIABILITY ANALYSIS

Relatore: Chiar.mo
Maurizio Bevilacqua

Tesi di Laurea di:
ALESSANDRO TORRESI

Correlatore:
Sara Antomarioni

A.A. 2019/2020

INDICE

INTRODUZIONE

Nel corso dell'ultimo secolo si è riscontrato un'impennata nella curva del progresso, basta pensare che si è passati dal periodo dell'industria 2.0, rivoluzione del web e nascita dei primi social network, avvenuta intorno ai primi anni del 2000, all'era dell'industria 4.0. Il termine venne usato per la prima volta nel 2013 e questa "tendenza" è caratterizzata dall'automazione industriale, presentando come punto di forza un lavoro di cooperazione tra uomo-macchina, che è stato possibile grazie all'introduzione di nuove tecnologie che hanno migliorato le condizioni di lavoro.

L'industria 4.0, dunque, viene vista come il quarto stadio di un percorso evolutivo dei sistemi produttivi, che parte dall'inizio della rivoluzione industriale ed arriva ad oggi; questo processo si fonda principalmente su 7 pilastri:

- *Simulation*: si prova a prevedere il comportamento di un sistema andando a creare un gemello digitale dello stesso, in modo tale da poter individuare quali sono le modalità di funzionamento;
- *Big data*: i sistemi altamente connessi producono una quantità enorme di informazione, perciò bisognerà effettuare un lavoro di elaborazione;
- Additive Manufacturing: la procedura che permetterà di sostituire o integrare i classici sistemi di manifattura sottrattiva, cercando di costruire il prodotto date le esigenze e in base alle caratteristiche che deve avere;

- *Autonomus Robot*: ovvero robot che possono eseguire specifiche attività, senza il controllo continuo dell'operatore;
- *Augmented Reality:* attraverso l'integrazione dei sistemi simulativi/virtuali, si cercherà di avere un duplicato/gemello digitale dell'impianto, per poter riuscire a prevedere i comportamenti;
- *Cloud Computing*: lo scopo sarà quello di svincolare le capacità di calcolo dei singoli operatori, passando alla capacità di calcolo diffusa (mappatura genoma umano)
- *Cybersecurity*: ovvero proteggere/evitare che avvengano furti o attacchi, fatti in modo da creare uno svantaggio rispetto ai concorrenti;
- *Internet of Things*: si tratta di creare sistemi che in base alla possibilità di accedere alla rete saranno sempre collegati;
- *Sistem Integration*: mettere insieme i diversi tasselli appena descritti può portare ad un sistema ampliamente integrato, in cui tutte le diverse attività che vengono svolte possono comunicare tra loro ed interagire con l'operatore per suggerire nuove modalità di intervento o per proporre azioni di tipo correttivo.

Nella tesi verranno utilizzati i pilastri di "*Big data*" e "*Simulation*". Il primo lo useremo come tema centrale, il nostro scopo sarà proprio quello di partire da un dataset molto ampio ricavato dalla FMECA e tramite regole di associazione estrapolare le informazioni più importanti. Il secondo invece lo incontreremo nel

primo capitolo quando citeremo un'integrazione alla metodologia della FMECA, la quale interesserà un'area chiamata ottimizzazione delle operazioni.

La continua ricerca dell'innovazione ci ha portato a richiedere sempre più elevate comodità per migliorare il nostro stile di vita, entrando in un sistema dove non è presente una vera e propria fine, ma semplicemente un continuo desiderio nel migliorare ciò che ci circonda.

In un ambito industriale, si vedrà come le crescenti capacità e funzionalità per molti prodotti stanno creando un livello superiore di complessità per il produttore, che sarà quindi tenuto a mantenere la qualità e l'affidabilità del servizio richiesto dai clienti. Ogni componente dovrà essere studiato al fine di valutare la migliore soluzione di manutenzione, in funzione del suo tasso di guasto, del suo costo e dell'impatto di difetto sull'intero sistema. In altre parole, per ogni macchina, è necessario decidere se è meglio attendere che avvenga il danno o se è maggior conveniente prevenirlo.

In quest'ultimo caso, il personale addetto alla manutenzione deve valutare se è meglio effettuare controlli periodici o utilizzare un'analisi progressiva delle condizioni di funzionamento. Chiaramente, un buon programma di manutenzione deve definire strategie diverse per le diverse strutture. È quindi necessario assegnare una priorità diversa ad ogni componente o macchina dell'impianto e concentrare gli sforzi economici e tecnici sulle aree che possono produrre i migliori risultati.

Lo scopo della tesi è determinare le regole di associazione per determinare eventuali problematiche, basandosi su un dataset relativo alla FMECA.

CAPITOLO 1

ANALISI DEL RISCHIO: FMEA/FMECA

In questo capitolo, verranno trattati i metodi di analisi del rischio, andando a presentare le componenti della FMECA (la FMEA e la CA), le loro caratteristiche e l'integrazione della simulazione Montecarlo per cercare di risolvere alcuni problemi relativi alla FMECA.

1.1 FAILURE MODE, EFFECTS AND CRITICALITY ANALYSIS

La FMECA, acronimo di *Failure Mode Effect and Criticality Analysis*, è un'analisi che deriva dal risultato di due fasi:

- Analisi delle modalità di guasto e degli effetti (FMEA);
- Analisi della criticità (CA).

La metodologia FMECA mette in evidenza le probabilità che un guasto si presenti insiemealle modalità con le quali ciò potrebbe accadere: quest'analisi viene svolta basandosi sul numero di priorità di rischio (RPN).

Lo scopo di questo studio minuzioso è quello di mettere in evidenza i punti deboli di un progetto in corso o ancora da iniziare e successivamente, se trovati “errori”, intervenire con adeguate modifiche in modo da eliminarli.

1.1.1 FAILURE MODE, EFFECTS ANALYSIS

La FMEA, acronimo di *Failure Mode and Effect Analysis*, è una tecnica di valutazione di possibili problemi di affidabilità durante le prime ore del ciclo di avanzamento: viene svolta durante questo arco temporale poiché è più semplice individuare i problemi che possono essere superati tempestivamente; così facendo si migliora la coerenza attraverso la progettazione.

La FMEA può essere applicata anche come metodologia di previsione per analizzare le modalità di guasto o di difetto di un elemento, in modo da poter ritrovare le cause di questi e soprattutto per valutare gli effetti che ne possono derivare.La metodologia, per determinare le conseguenze, dovrà analizzare i dati raccolti dalle analisi dei guasti passati, ma dobbiamo tenere in considerazione che trattandosi di previsioni non si avrà mai una precisionedel 100%; ciò comporterà nuovamente l’utilizzo di questa tecnica (o di altre tecniche) con cadenze precise ed infatti, molto spesso,può essere paragonata ad un algoritmo ricorsivo, poiché come questo, cercherà sempre di migliorarsi.

Per il calcolo del rischio del metodo FMEA possiamo descrivere i 3 componenti principali che vanno ad incidere:

- la gravità(S);
- l'evento(O);
- il rilevamento(D).

Questi elementi dovranno successivamente essere moltiplicati tra loro per determinare il numero di priorità di rischio (RPN). Le tre componenti sono indicate come variabili numeriche e devono essere comprese fra l'1 e il 10 (estremi compresi), la loro valutazione viene posta in base ad una scala crescente; ciò ovviamente comporterà che l'indicatore RPN potrà avere un valore compreso tra l'1 e 1000.

	Gravità (S)	**Evento (O)**	**Rilevamento (D)**	**RPN=S*O*D**
Guasto potenziale 1	3	10	6	180
Guasto potenziale 2	10	3	6	180
Guasto potenziale 3	3	6	10	180
Guasto potenziale 4	10	6	3	180

Tabella 1Esempio del calcolo di rischio (FMEA)

Per poter effettuare un'analisi adeguata bisognerà valutare gli elementi in base alla loro importanza e al valore che posseggono.

Prendendo in esame la Tabella 1, per determinare un ordine di priorità partiremo analizzando la *gravità*, i guasti 2 e 4 presentano entrambi il valore massimo possibile, mentre le rimanenti presentano un valore pari a 3. La seconda componente da tenere in considerazione è l'*evento*, prendendo in esame delle coppie tra il guasto potenziale 2 e il 4, quello che presenta il valore maggiore (6) è il secondo, mentre nella coppia guasto potenziale 1 e 3 si vede che è il primo guasto che presenta un valore maggiore (10). Perciò l'ordine di priorità che troviamo è: 4-2-1-3.

La stima dei pesi è una fase cruciale nel calcolo dell'indice di criticità, per questo molto spesso accade che una volta imposti, questi vengano messi in discussione, poiché la possibilità di incorrere in errori è elevata e la possibilità di una valutazione errata dei coefficienti è tanto più alta tanto maggiore è il numero di persone coinvolte nell'analisi. Per questo motivo tali metodi di analisi permettonoagli analisti di correggere a posteriori le stime effettuate sui coefficienti, così da poterle sostituire con valori più precisi.

1.1.2 CRITICALITY ANALYSIS

La FMECA rappresenta un'estensione della FMEA, poiché in aggiunta viene inclusa l'analisi di criticità (CA).Questo è uno studio che permetterà di aiutare

nell'analisi del processo di produzione o di assemblaggio e di documentare le cause dei cambiamenti.

Possiamo dividere l'analisi di criticità in due tipologie:

1. quantitativo,che serve per determinarsi il valore della criticità tramite una serie di passaggi:
    1) definire l'affidabilità/inaffidabilità per ogni elemento, in un dato momento operativo;
    2) identificare la parte di inaffidabilità degli elementi e valutarne la probabilità di perdita (o la gravità) che risulterà da ogni modalità di guasto che può verificarsi;
    3) calcolare la criticità per ogni potenziale modalità di guasto;
    4) calcolare la criticità per ogni voce ottenendo la somma delle criticità per ogni modalità di guasto.
2. qualitativo, che servirà per valutare il rischio e dare priorità alle azioni correttive e si determinerà svolgendo una serie di passaggi:
    1) valutare la gravità dei potenziali effetti di un guasto;
    2) valutare la probabilità che si verifichi un guasto per ogni modalità potenziale;

3) confrontare le modalità di guasto tramite una matrice di criticità, che identifica l'occorrenza sull'asse orizzontale e la gravità sull'asse verticale.

1.2 L'INTEGRAZIONE DELLA SIMULAZIONE MONTECARLO

Una strategia di manutenzione inadeguata può comportare perdite notevoli e situazioni imprevedibili: è infatti importante riuscire a determinare un adeguato sistema o quanto meno una strategia di manutenzione per evitare di incorrere in situazioni spiacevoli (esempio: blocco di un settore data la rottura di un macchinario, danno ad un dipendente).

Possiamo quindi vedere la tecnica FMECA come un'applicazione usata per determinare l'affidabilità del processo, tenendo in considerazione le cause potenziali dei guasti e i loro effetti sul sistema in esame.

Ciò però non sarà sempre possibile data la possibilità di incorrere in errori causati da un'assegnazione errata dei pesi durante lo studio dell'MPN: sarà dunque necessaria l'applicazione di metodi euristici. Un esempio di approccio potrebbe

essere presentato dalla combinazione del "confronto fra coppie" e dalla "simulazione Montecarlo".

Si inizia prendendo in considerazione i pesi che uno o più decision maker attribuiranno: infatti, essendo spesso fonte di errore, dovranno essere eliminati; e si dovrà stilare una classifica ordinata per rango di importanza.

Verrà poi adottato un confronto a coppie per determinare quale sarà più importante attribuendo 3 diversi possibili giudizi:

- più importante con punteggio 2;
- meno importante con punteggio ½;
- indifferente con punteggio 1.

Si definirà così la matrice di giudizio, la quale servirà per calcolare un vettore di priorità per ponderare gli elementi della matrice. La parte relativa al confronto tra coppie termina calcolando i componenti normalizzati del vettore autogeno destro della matrice finale, ricavando così la priorità finale.

La parte della simulazione Montecarlo invece, consiste nel conferire pesi casuali alle componenti del vettore di priorità finale, in modo da poter esplorare in modo efficiente i risultati di molte combinazioni di pesi.

Il vantaggio di questo strumento è la possibilità di analizzare in modo più dettagliato la robustezza della classifica e di tutte le alternative per poter ricavarci

l’opzione migliore (grazie alla possibilità di modifica in contemporanea di tutti i pesi).

CAPITOLO 2

REGOLE DI ASSOCIAZIONE

In questo capitolo, verranno trattate le regole di associazione: si parte da una spiegazione del data mining, per poi proseguire con la storia, la metodologia e le applicazioni delle regole di associazione.

2.1 DATA MINING

Il data mining è un sottocampo interdisciplinare di informatica e statistica con la funzione di estrarre informazioni (attraverso l'applicazione di metodi intelligenti) da una grande quantità di dati (es. banche dati, datawarehouse, ecc.) e di trasformali in una struttura differente e soprattutto comprensibile per un altro eventuale utilizzo.

Per determinare le regole dovremo effettuare un lavoro di estrazione: il data mining permette tale lavoro tramite l'applicazione del metodo di scoperta.

Si conoscono due metodi che guidano le applicazioni del data mining: quello appena citato, il quale in maniera automatica scopre importanti relazioni tra i dati, per poi setacciarli alla ricerca di vere e proprie regole (similitudini, tendenza, ecc..), l'altro è il metodo di verifica, il quale a differenza del precedente non crea nessuna

nuova informazione, ma bensì studia un'ipotesi posta dall'utente e ne verifica la validità nei dati.

In base alle esigenze di utilizzo, si necessiterà di alcune trasformazioni da parte dei dati, sarà dunque necessaria l'applicazione di una o più regole, quali:

- Normalizzazioni: ossia calcoli basati sulla distanza tra punti nello spazio multidimensionale (ed in quanto la distanza è una misura, si necessiterà di valore positivo/normalizzato);
- Smoothing: è il livellamento dei valori per renderli più omogenei per evitare di arrivare ad un degrado del risultato finale (un esempio è la media mobile o l'arrotondamento);
- Trattamento dei dati mancanti: è possibile incorrere in dati mancanti nei dataset e perciò tramite tecniche secondarie si cerca di risolvere la problematica; ad esempio si possono creare valori con cui si può andare a sostituire il dato mancante, sarà dunque opportuno, andare a confrontare il risultato con modelli creati precedentemente che siano in possesso di tutti i dati, in quanto altrimenti si potrebbe insorgere in analisi falsate.

Possiamo concludere nel dire che il Data Mining è un processo fondato sulla ricerca di pattern utili, più che un algoritmo per la risoluzione di dati, e si fonda principalmente su tre macro-processi: esplorazione, modellazione e valutazione [Fig1].

Figura 1Modello di processo del Data Mining

[http://tesi.luiss.it/17593/1/078262_FUGGITTI_FILIPPO.pdf]

La fase di esplorazione è rappresentata dai dati su cui si poggia il processo, al fine di prendere decisioni più precise, questa procedura mette le basi per la fase di modellazione, la quale necessiterà di informazioni quanto più precise possibili, per la costruzione di modelli previsionali, ed infine una volta preso in esame il modello lo si dovrà paragonare ad altri dove si conoscono già i risultati, stiamo dunque parlando della fase di valutazione. Ovviamente in tutto questo si sta lavorando con sistemi previsionali e dunque non precisi, infatti i risultati effettivi, in genere,

risultano inferiori: sarà richiesto dunque di rimettere mano sulla fase di modellazione, per poter avere un tipo di risposta più accurata.

2.2 STORIA

Il concetto di regole di associazione è stato reso popolare in particolare dall'articolo del 1993, dove vengono utilizzate per l'analisi degli acquisti all'interno dei supermercati, di Rakesh Agrawal, ed è stato introdotto sempre da Agrawal insieme a Tomasz Imielinski e Arun Swami. Tuttavia si possono ritrovare anche nel documento del 1966 su GUHA, un metodo di data mining generale sviluppato da Petr Hàjek et al, dove considerati due sistemi finiti, uno non vuoto di oggetti ed uno di proprietà, e dove si conosce se un elemento del primo sistema possegga o meno un determinato elemento del secondo, si può arrivare a definire, grazie alla ricerca di relazioni, quali proprietà valgono per tutti o quasi per tutti gli oggetti; questo metodo dunque può essere visto come una sostituzione di un'intuizione, in una certa fase della ricerca scientifica.

Nel 2015 sono diventate argomento tra i più citati nel data mining in quanto ha acquisito più di 18000 citazioni secondo Google Schoolar.

2.3 TEORIA

Le regole di associazione sono informazioni derivanti da relazioni fra i dati presenti in un database, tali informazioni rappresentano i comportamenti/ le probabilità e si presentano dunque come dati di previsione statistica. Possiamo dividere le informazioni di questo processo in due parti; quella antecedente, che presenta i dati già presenti nel database, e una conseguente, la quale è un elemento trovato come combinazione degli elementi della parte antecedente.

Le regole presentano determinate proprietà con lo scopo di evidenziare le relazioni appropriate allo studio che si necessita e vengono determinate grazie all'utilizzo di appositi algoritmi. Infine per l'estrazione di queste, possiamo citare l'utilizzo del data mining, un insieme di tecniche e metodologie con lo scopo di scovare eventuali relazioni tra i dati.

Le association rules, sono dichiarazioni "if-then" ed hanno come funzione quella di mostrare la probabilità delle relazioni tra i dati (i quali possono essere più o meno simili), all'interno di grandi dataset, in vari tipi di database, con forma

$X \rightarrow Y$ (dove X implica Y).

Le regole che verranno estratte, possono presentare caratteristiche o funzioni differenti, in base a quanto saranno performanti ed alle necessità della relativa applicazione sulla quale verrà svolto il lavoro.

Un esempio di questa eterogeneità potrebbe essere la mancanza di sicurezza di una regola, se pensiamo ad un sistema di ambito medico, la certezza della diagnosi dovrà essere assoluta, senza la possibilità di mostrare margine di errore. Oppure, cambiando ambito applicativo, si potrebbe pensare ad un sistema dove si studia il rapporto tra le vendite di un certo prodotto e quelle totali: in questo caso si potrebbe incorrere nel non considerare le vendite di determinati articoli venduti in quantità minore ma molto redditizi. Nella situazione tipica in cui nel magazzino siano presenti numerosi articoli di tipo diverso, si riscontra che solo una piccola percentuale degli stessi è responsabile di un elevato valore percentuale del fatturato ed infatti studiando la gestione selettiva delle scorte basata sulla tecnica ABC, capiamo che gli articoli C rappresentano il 70% circa delle quantità presenti nel magazzino, e verranno dunque venduti di più, rispetto agli altri articoli presenti, che ricoprono solo il 10/15% del profitto all'azienda, mentre se si considera quelli di livello A, i quali occupano il 10/15%, noteremo un profitto del 75% circa, si rischierebbe dunque di non considerare la parte più remunerativa per l'azienda.

Per ovviare a queste problematiche sono presenti due metriche (caratteristiche), con valori compresi tra 0 e 1, per misurare l'utilità del risultato:

- Support (supporto): ovvero l'indicatore della frequenza con cui gli elementi appaiono nei dati:

$$S = P(X, Y) = \sigma(X, Y) / N \quad \text{con } X \rightarrow Y;$$

- Confidence (confidenza): ovvero l'indicatore del numero di volte in cui le affermazioni if-then vengono trovate vere:

$$C = P(Y|X) = \sigma(X, Y) / \sigma(X) \quad \text{con } X \rightarrow Y.$$

Una buona regola dovrà presentare un buon supporto e una buona confidenza, senza eventuali eccessi, questo perché prendendo un margine di supporto troppo alto, si rischia di incorrere nel problema ABC sopra citato, nel caso opposto invece si rischierebbe di prendere regole non troppo performanti, con costi significativi; invece per la confidenza si potrebbe non avere più la tolleranza necessaria, tralasciando dunque valori importanti.

2.4 GRAFI

Le relazioni sono rappresentabili visivamente da strutture denominate grafi

G = (V, A), le quali sono costituite da un insieme di vertici/nodi V, che dovranno definire i dati antecedenti, e da un insieme di archi/spigoli A, che rappresenteranno invece i dati conseguenti (ovvero le regole).

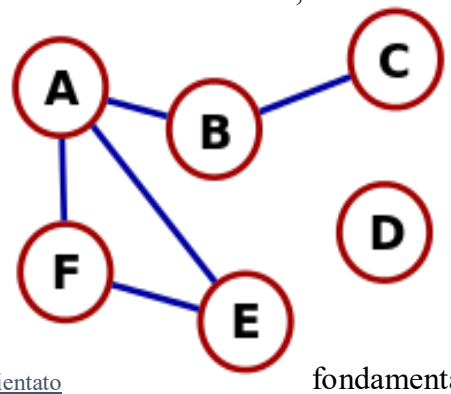

Figura 2 Grafo non orientato [https://it.wikipedia.org/wiki/Grafo]

Dividiamo principalmente i grafi in due categorie fondamentali,

quelli non orientati [Fig.2] e quelli orientati [Fig3].

La differenza sostanzialmente viene sottolineata dal nome, i grafi orientati hanno una direzione e indicano un collegamento unidirezionale da un nodo all'altro, e proprio grazie a questa caratteristica di unidirezionalità che si possono utilizzare come rappresentazione grafica delle regole.

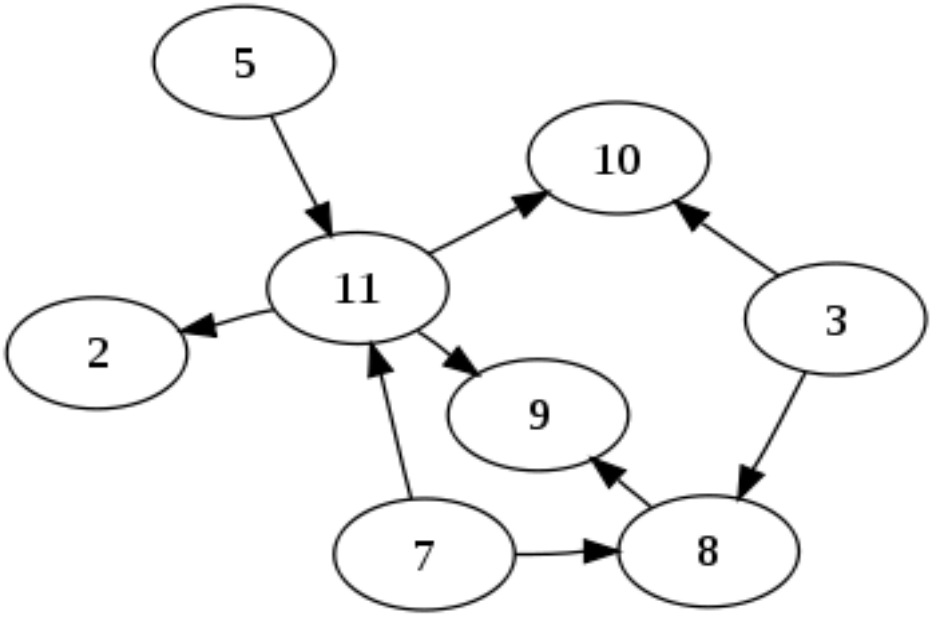

Figura 3 Grafo (aciclico) orientato

[https://it.wikipedia.org/wiki/Digrafo_aciclico]

Un esempio di applicazione di grafo è l' "Albero" [Fig. 4], il quale rappresenta una tipologia facente parte dei gruppi non orientati, presentando due vertici connessi da

un unico cammino, ma possedendo un'organizzazione dati basata su di una determinata gerarchia.

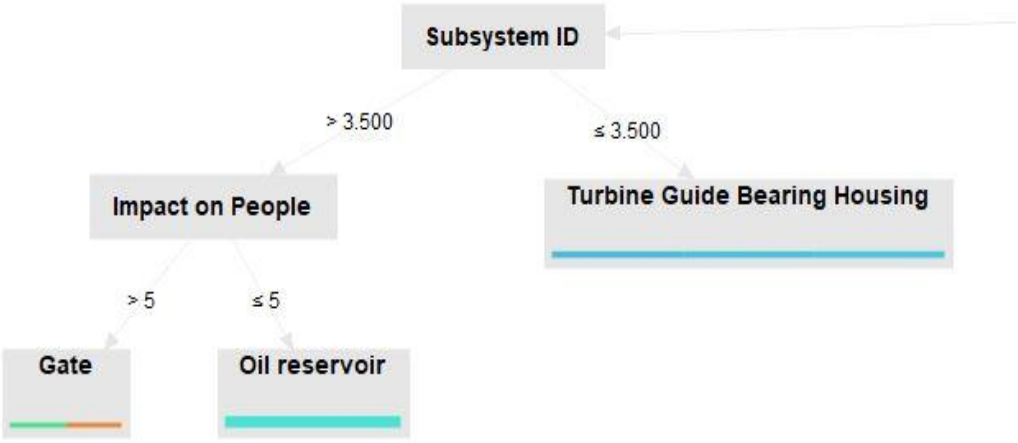

Figura 4Parte di un grafo ad albero

2.5 ALGORITMI

2.5.1 Apriori

All'interno di un database si possono trovare quantità elevate di informazioni, può accadere che si ci siano un numero di dati frequenti alto, perciò esistono varie strategie per determinare questi itemset, ovvero:

- ridurre il numero di candidati;
- ridurre il numero delle transazioni;
- ridurre il numero delle comparazioni.

Nella strategia di riduzione del numero di candidati, troviamo l'algoritmo Apriori. Il principio Apriori dice che se un itemset è frequente, allora anche tutti i suoi sottoinsiemi devono esserlo e si basa sulla proprietà del supporto "anti-monotona", ovvero:

$$\forall X, Y (: X \subseteq Y) \Rightarrow s(X) \geq s(Y);$$

questa formula indica che il supporto di un itemset non eccede il supporto dei suoi sottoinsiemi.

L'algoritmo si divide in due funzioni consequenziali, la prima parte analizzerà tutti gli elementi possibili ed eliminerà tutti quelli che hanno una soglia di supporto inferiore alla soglia prefissata, mentre la seconda funzione genererà K+1-itemset da quelli rimasti, sfruttando la proprietà anti-monotona, si applicherà infine un lavoro di check up, dove se si manterrà nuovamente la soglia del supporto si potrà andare avanti per ripetere la procedura e così via.

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3

Itemset	Count
{Bread, Milk}	3
{Bread, Beer}	2
{Bread, Diaper}	3
{Milk, Beer}	2

Itemset	Count
{Bread, Diaper, Milk}	3

Diaper	4	{Milk, Diaper}	3
Eggs	1	{Beer, Diaper}	3

Tabella 2 Visualizzazione del principio Apriori

Nell'esempio [Tab.2] abbiamo lo sviluppo del principio Apriori, rappresentato da una tabella che da sinistra verso destra avrà un'evoluzione, presentando, di step in step, un numero di itemset sempre maggiore (partendo da 1). Durante questa evoluzione si dovrà tener conto del vincolo di maggiore o uguale del "valore minimo del supporto" pari a 3. Partendo dalla prima tabella, notiamo che la colonna del count presenta 2 valori che non soddisfano le richieste, quelle riferite alla coca e alle uova, perciò non sarà necessario generare i candidati che le coinvolgono. Arrivati alla tabella due applicheremo lo stesso ragionamento, per quei valori che saranno sempre minori o uguali a 3, arrivando infine alla tabella 3, che ci darà la nostra soluzione.

2.5.2 FP-Growth

Nell' algoritmo Apriori troviamo però delle carenze, ovvero la generazione di oggetti candidati, i quali sono necessari, ma potrebbero diventare molto numerosi, se la presenza di itemset nel database è elevata, e la necessità di scansioni multiple

della memoria, per controllare il supporto di ogni itemset generato, portando dunque ad un costo elevato ed inutile.

Per far fronte a queste carenze è stato introdotto l'FP-Growth, le lettere iniziali che compongono la sigla di questo algoritmo indicano "frequent pattern" e dunque, come suggerisce il nome, genera automaticamente un pattern frequente, senza la necessità di generare candidati, risolvendo dunque questo primo problema dell'Apriori.

L'esecuzione dell'FP-Growth si riassume principalmente in due passaggi: quello iniziale di analisi e di memorizzazione, dove prima conta le occorrenze degli elementi (visualizza l'attributo associandolo al rispettivo valore) nelle transazioni, per poi memorizzare il tutto in una tabella di intestazione, e quello finale di costruzione, dove immetterà i dati presenti nella tabella, in un grafo ad albero (FP-Tree).

Gli elementi presenti vengono rappresentati tramite un ordine decrescente della loro frequenza, rendendo l'albero molto chiaro a livello di lettura ed ovviamente ponendo un valore minimo di supporto, avverrà un singolo e rapido lavoro di potatura per quegli elementi che non soddisferanno il vincolo, ottimizzando i costi. Inoltre una caratteristica molto importante presente in questa struttura dati, è la facilità di aggiornamento del sistema nel momento in cui si ha intenzione di

aggiungere nuove transazioni, non richiedendo ulteriori accessi al database e senza il bisogno di ulteriori verifiche dato il lavoro di taglio del supporto sopra citato.

2.6 APPLICAZIONI

Attualmente, i campi di applicazione sono molteplici e comprendono: diagnosi mediche, censimenti, utilizzo del web, relazione con i clienti ed analisi degli acquisti all'interno dei supermercati.

Per illustrare meglio questo concetto si andrà a presentare come esempio l'ultimo citato (in quanto è l'applicazione sulla quale nascono le regole associative).

Studiando gli scontrini emessi dai registratori di cassa possiamo analizzare il comportamento degli acquirenti medi e fare uno studio a riguardo compilando uno schema [Tab.3].

	Milk	Bread	Tomato	Egg	Coke	Meat
1	1	1	0	0	0	0
2	1	1	1	1	0	1
3	0	1	1	0	1	1
4	1	1	1	0	0	1
5	1	1	1	0	1	0

Tabella 3Esempio di base di dati con 6 oggetti e 5 transazioni

L'analisi avverrà su di una tabella a doppia entrata dove le colonne indicano il prodotto di vendita, mentre le righe i clienti. Questa matrice presenta all'interno una variabile binaria dove con 1 si indica la presenza del prodotto nello scontrino del cliente, mentre con 0 la sua assenza (si poteva applicare anche con una relazione Vero/Falso al posto della variabile).

Leggendo la tabella, risulta subito semplice capire cosa hanno comprato i nostri clienti e saltano all'occhio anche alcune regole, come ad esempio:

- {Milk} → {Bread};
- {Tomato, Meat} → {Bread}.
- {Bread, Meat} → {Tomato};

Prendendo il terzo caso in esame, chi ha sia il pane che la carne nella sua lista della spesa, avrà molto probabilmente anche il pomodoro (il viceversa non è accettato): potenzialmente questo avviene perché vorrà preparare un hamburger, ciò ci dà come segnale quello di porre il più vicino possibile questi 3 alimenti negli scaffali, per permettere al cliente di acquistare subito i nostri prodotti.

Questo trattandosi ovviamente di un esempio esplicativo, possiede un numero di dati non sufficientemente soddisfacente per uno studio di questo tipo, si dovrà dunque avere un aumento di informazioni, le quali provocheranno una maggior complessità, che a sua volta richiederà l'introduzione di nuove metodologie per la risoluzione.

Tra le problematiche che riscontreremo infatti troveremo sia i tempi, che si allungheranno inevitabilmente, sia la presenza di numerose regole, che sono ciò di cui necessitiamo, ma esse rappresenteranno un ostacolo per la navigazione e la ricerca di quelle più interessanti. Per poter ottimizzare le risorse, perciò, si applicheranno o dei linguaggi di programmazione oppure programmi per semplificare il lavoro tramite codice, come ad esempio RapidMiner.

CAPITOLO 3

SOCIAL NETWORK ANALYSIS

3.1 L'ANALISI DELLA RETE SOCIALE

L'analisi della rete sociale è una metodologia di analisi delle relazioni interpersonali. Nasce per studiare le relazioni che intercorrono tra gli individui e determinarne le reti sociali; di recente questa tecnica ha trovato applicazione in molti altri ambiti diversi da quello sociale, nonostante abbia mantenuto l'appellativo "social", come ad esempio l'ambito fisico, biochimico e genetico.

Il sistema sul quale la SNA (Social Network Analysis) si basa è una società abitata da persone, che prendono il nome di attori, le quali, tramite interazioni con gli altri abitanti, plasmeranno sia il loro carattere sia quello delle altre persone a loro vicine. Nella fase di sviluppo della metodologia e per la sua rappresentazione è stato necessario introdurre strumenti e concetti derivanti dalla teoria dei grafi, ossia la disciplina che studia oggetti discreti che permettono di sintetizzare in uno schema una molteplicità di situazioni e processi, e che consente di analizzare in termini quantitativi e algoritmici i grafi.

Nel grafico derivante dalla tecnica sopra descritta, verranno rappresentati gli attori, che saranno i nodi, e le relazioni sociali che saranno gli archi.

Esistono differenti tipologie di grafi, ma la tipologia che viene spesso usata per questo studio è quella dei grafi firmati, ovvero grafi specifici dove viene posto un segno positivo o negativo vicino all'arco per indicare se gli individui tra loro sono in buoni rapporti, nel primo caso, o se in cattivi rapporti nel secondo caso.

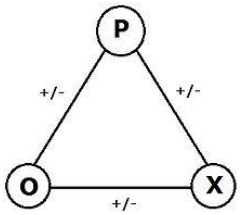

Figura 5 Grafico firmato

Per la varietà di grafi sopra descritta esisto due diverse varianti, che si distinguono per il tipo di ciclo presente:

- equilibrato: indica un ciclo dove il prodotto di tutti i segni è positivo;
- sbilanciato: indica invece il prodotto negativo.

La Teoria dell'equilibrio afferma che: gli attori, i quali prendono parte ad un ciclo equilibrato, saranno meno soggetti a cambiare opinione sul loro legame, mentre quelli presenti ad un ciclo squilibrato, avendo un'opinione mutevole sui loro legami, tenderanno a distaccarsi e coloro che resteranno andranno a formare un ciclo di tipo equilibrato. Questa tesi sfrutta entrambe le tipologie di cicli, perciò si potranno trovare delle regole forti e più utili alle applicazioni.

Dopo aver classificato alcune tipologie di grafo, per comprendere al meglio le applicazioni della social network analysis, è necessario introdurre il concetto di metriche relative alla SNA, ne esistono diverse tipologie, ma ne andremo a prendere in esame soltanto due:

– Out-Degree (OD), la somma ponderata degli *n* bordi in uscita dal nodo *j*;

$$|A| = \sum_{v \in V} \deg^+(v)$$

Figura 6Formula dell'Out-Degree

La OD indica l'influenza che un nodo j avrà sui nodi che lo succedono; tanto più il valore presentato sarà alto, tanto più il nodo preso in esame sarà importante: questo si tradurrà in una necessità di effettuare dei controlli.

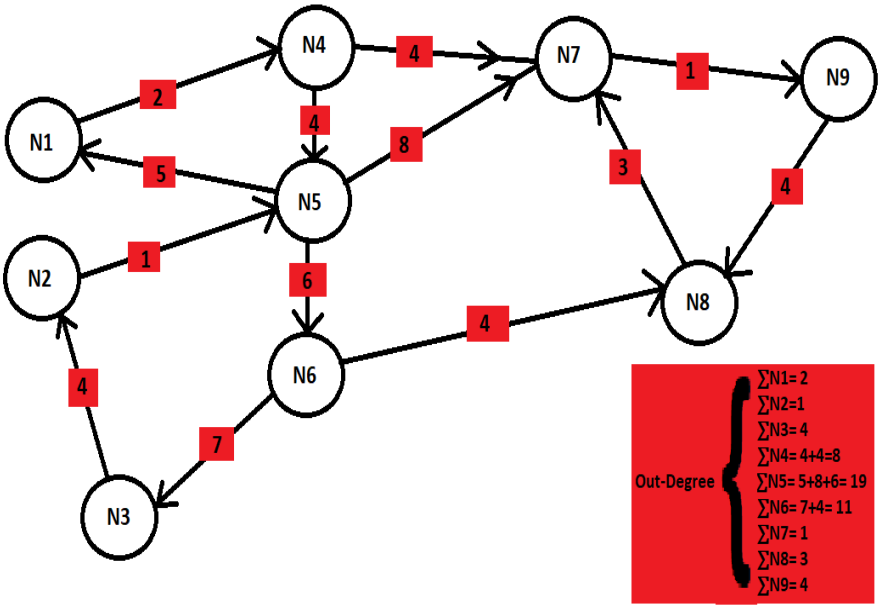

Figura 7Rappresentazione di un Out-Degree

Osservando la figura 3, notiamo come i valori dell'OD ci indicano che l'attore, raffigurato come il nodo N5, sarà quello con un peso maggiore in uscita, mostrandosi dunque come elemento con maggior influenza nel grafico.

- BetweennessCentrality(BC), la somma dei percorsi ponderati più brevi su cui appare il nodo *j*.

La metrica BC, invece, indica l'influenza che un nodo avrà sul grafo: se presenta un valore elevato potrà essere considerato come un ponte tra parti separate del grafo, permettendoci di arrivare in zone altrimenti inaccessibili. Per il calcolo di questa metrica partiamo dal considerare che per ogni coppia di vertici (s, t), si dovranno inizialmente calcolare i percorsi più brevi tra di loro, successivamente bisognerà determinarsi la frazione di commini più brevi che passano per un vertice v, ed infine sommare tali frazioni su tutte le coppie di vertici(s, t). In modo più compatto possiamo scrivere la formula come:

$$BC(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Figura 8Formula della BC

[https://dl.acm.org/doi/pdf/10.1145/3230485]

Nella figura 9, applicando la BC, si può concludere che ci sono nodi in cui si passa più frequentemente rispetto ad altri: infatti il nodo 3 è quello più influente nel grafo

avendo un valore di 15, mentre altri nodi come 1,7,8 hanno un valore pari a 0, risultando dunque inutili a questo studio.

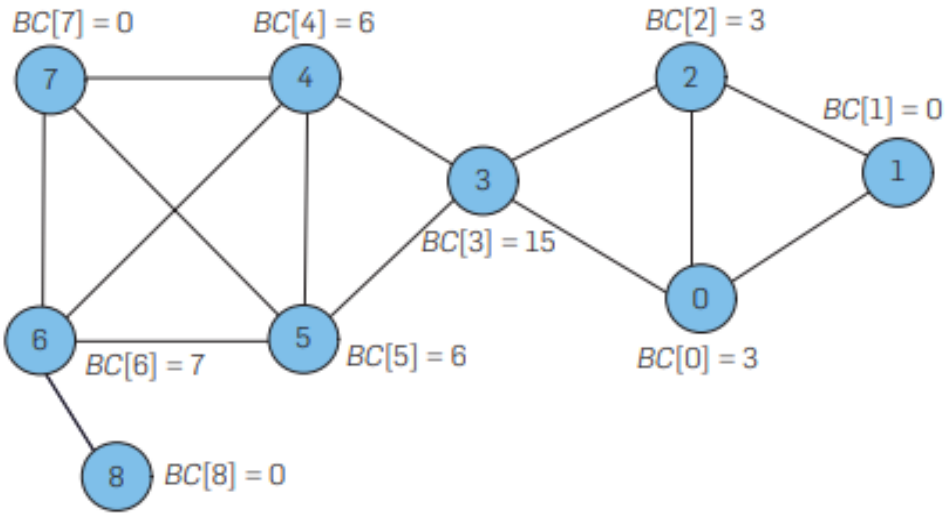

Figura 9 Descrizione di una BetweennessCentrality [https://dl.acm.org/doi/pdf/10.1145/3230485]

È importante parlare anche delle rappresentazioni visive, poiché risulta più vantaggioso raffigurare i nodi dei grafi attribuendogli colori, dimensioni ed altre proprietà avanzate in modo da assumere maggior rilievo: lo scopo è di trasmettere informazioni complesse; è necessario prestare un elevata attenzione all'interpretazione delle proprietà, poiché una sbagliata applicazione potrebbe portare a rappresentare il grafico in modo errato.

Possiamo concludere dicendo che la Social Network Analysis è una metodologia che svolge due funzioni molto importanti:

1. svolge un lavoro di rappresentazione grafica attraverso i grafi delle regole di associazione;
2. individuazione degli attori più critici: è dunque possibile stabilire quali attori non saranno utili al nostro studio e quelli che richiederanno una maggiore attenzione in quanto sono più importanti.

3.2 APPLICAZIONE DELLA SOCIAL NETWORK ANALYSIS

Sempre rimanendo nel tema del sociale, poniamo l'esempio dei social network, l'insieme delle persone alle quali un individuo è collegato ai giorni nostri può essere facilmente collegabile proprio tramite queste piattaforme e le informazioni potranno, ovviamente, poi essere sfruttate in vari contesti.

I vicini di rete di una persona possono presentarsi con relazioni profondamente diverse in base al tipo di rapporto, possiamo infatti trovare membri della famiglia, amici, colleghi, conoscenti ed altre tipologie di persone. Uno degli elementi fondamentali, nei social network, è proprio riuscire a distinguere le relazioni per poterle dare un adeguato peso, questa procedura prende il nome di "forza di legame", essa si riferisce alla "vicinanza" dell'amicizia ed andrà da legami più forti, come per i parenti, a quelli più deboli, come per magari i semplici conoscenti.

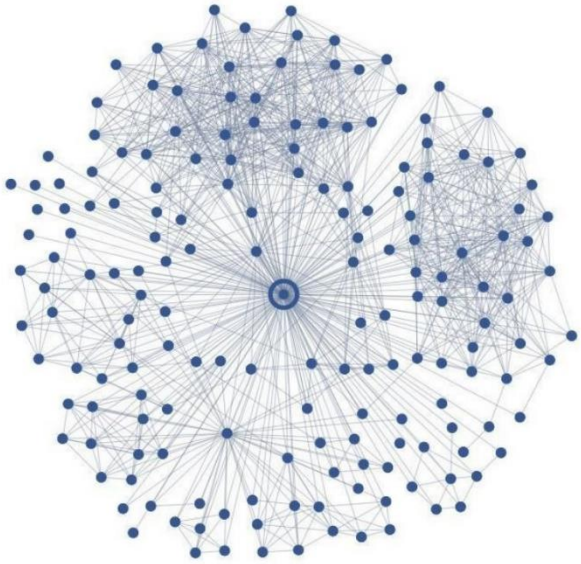

Figura 10 Un quartiere di Facebook

[https://dl.acm.org/doi/pdf/10.1145/2531602.2531642]

[Figura 5] Prendiamo ora in esame un utente facebook dichiaratamente sposato, si voglia determinare il proprio coniuge senza che venga rivelato il nome. Iniziamo studiando il problema, sappiamo che una relazione intima gioca un ruolo importante per i legami sociali della persona, è importante considerare che la relazione avviene tra due individui e che tra di essi avverrà un legame forte.

Una volta considerati i dati cerchiamo di rimpicciolire il cerchio ponendo delle fasce di età, ad esempio l'individuo da determinare dovrà avere almeno 20 anni, e che abbia un intervallo di amici che vada tra i 50 e i 2000, il limite inferiore serve

per eliminare gli utenti inattivi, se il coniuge sarà inattivo potrebbe non aver aggiornato il proprio profilo falsando la ricerca, mentre il limite superiore ha la funzione di eliminare le persone "famose", quest'ultime inseriranno un'ulteriore categoria, ovvero i fan, le quali non avranno alcun legame con i coniugi e quindi presenteranno delle relazioni inutili al problema.

La caratteristica di base che dobbiamo considerare è che ovviamente in quanto sposati dovranno essere amici, dunque è praticamente certo che ci sarà un qualche numero di amici condivisi, perciò basandosi su questa condizione valutiamo gli approcci:

- EMBEDDEDNESS (integrazione) si considera che probabilmente i coniugi avranno in comune un elevato numero di amici in comune, si ricerca dunque questa relazione di base;
- DISPERSION (dispersione) si considera che i coniugi avranno, ovviamente, degli amici in comune, ma sarà molto probabile che i due fungeranno più che altro da ponte fra gruppi di persone, più che essere entrambi presenti in un focolaio di persone molto unite, in quanto ognuno avrà avuto la propria vita e il proprio percorso.

Si stima che il primo approccio dia una precisione del 25%, mentre il secondo più del doppio, arrivando a circa il 60% della precisione, questo è dovuto a delle considerazioni importanti attuate dalla dispersione, prima tra tutti: il tipo di legame.

Ogni coniuge avrà avuto una vita separata dall'altra, unica, solo quella determinata persona avrà precisamente quel tipo di relazione con quelle precise persone. Già qui, riusciamo a vedere le prime differenze tra i due sposi, entrando più nello specifico possiamo approfondire dicendo che non tutte le persone sono uguali, avremo i famigliari, gli amici, i colleghi di lavoro, gli ex-compagni di scuola ed infine anche gli sconosciuti, questo implica che le relazioni sono molto differenti a livello di importanza, passando da un legame più forte, come per i famigliari, ad uno più debole, come per gli sconosciuti. Perciò possiamo dire che i focolai, si presenteranno differenti tra le due persone in questione, sebbene magari un marito abbia rapporto con parenti, colleghi ed amici della moglie, questo non implicherà che lui debba avere necessariamente lo stesso legame che la coniuge avrà con i suddetti, sarà anzi impossibile, ciò indicherà che i focolai non sono confinanti, dimostrando così l'enorme diversità con l'approccio dell'embeddedness. Possiamo dunque sintetizzare la struttura come i coniugi u e v, che rappresenteranno i nodi che collegheranno(ponte) due diversi focolai.

Possiamo anche confrontare queste misure con altre diverse forme di relazione, bisogna infatti tener conto della visualizzazione reciproca dei profili, l'invio dei messaggi, la co-presenza agli eventi e soprattutto la presenza di entrambe le persone su di una foto. Le migliori prestazione le troviamo proprio in questa azione, si studia infatti in un arco di tempo ridotto (es. 90giorni), quante volte si presenta questa

eventualità ed anche solo con questa opzione si riesce a trovare o il coniuge, o un famigliare o infine il migliore amico con una precisione del 60%.

Un'altra importante fonte di variazione per gli utenti è la dimensione dei loro quartieri e il tempo trascorso da quando si sono iscritti su facebook, queste due caratteristiche risultano essere correlate, è infatti logico comprendere che non appena ci si iscrive il quartiere di rete aumenta significativamente. Questa crescita presenta due potenziali effetti sul livello di performance, agendo in direzioni opposte, un quartiere di lavoro in rete più maturo presenterà una complessità maggiore, nuocendo alla performance, ma probabilmente rifletterà anche sulle relazioni off-line in modo più ricco di dettagli, migliorando in questo modo la performance

È possibile trovare un'evoluzione della dispersione, ovvero la dispersione ricorsiva, si tratta di un approccio dove una volta identificati i nodi v per i quali il collegamento u-v raggiunge una elevata dispersione normalizzata basata su un insieme di vicini comuni C uv che, a loro volta, hanno anch'essi un'elevata dispersione normalizzata nei loro collegamenti con u. La norma è una semplice combinazioni tra due valori che possono essere ricavati, ovvero il disp(u, v) e il emb(u, v), sono rispettivamente il valore ricavato dal metodo della dispersione e quello tramite il metodo dell'embeddedness. La formula di tale procedura è:

$$|N| = disp(u, v) / emb(u, v)$$

troviamo che le prestazioni sono più elevate quando aumenta la disp(u, v), essendo il numeratore, a svantaggio dell'emb(u, v), essendo il denominatore.

Possiamo dunque define la dispersione ricorsiva come il principio più performante per la risoluzione del nostro problema portando una percentuale di riuscita del 76,9%.

3.3 LA SOCIAL NETWORK ANALYSIS APPLICATA ALLA FMECA

Uno dei pilastri dell'industria 4.0 è rappresentato dai big data ovvero una raccolta molto grande di dati informativi, spesso eterogeni, che dovranno essere analizzati per riuscire a trovare dei collegamenti, i quali verranno infine estratti; questo è il lavoro che effettua per la FMECA.

Per la risoluzione di tale metodologia, si attua una combinazione tra due metodologie:

- l'Association Rule Mining, è un tipo di data mining che viene utilizzato per determinare le relazioni attributo-valore che si presentano con una frequenza in base alla probabilità del verificarsi di un guasto;
- la Social Network Analysis, la quale viene utilizzata per fornire una rappresentazione visiva delle reti costituite dalle regole di associazione, e

soprattutto per evidenziare gli eventi più critici, essendo la loro presenza un pericolo per l'intero impianto o anche più grave, per l'azienda stessa.

La procedura si basa sui dati raccolti dalle analisi dei fallimenti passati: è dunque fondamentale essere in possesso delle informazioni riguardo le criticità dell'azienda, in modo da poter lavorare senza pregiudizi di alcun tipo, potendo mettere in discussione tutte le prospettive così da ottenere una comprensione completa dei potenziali fallimenti e delle criticità dei guasti.

Verrà dunque creato un dataset focalizzato sulle apparecchiature prese in esame, dalle quali si evidenzieranno le potenzialità dei guasti, le loro criticità e gli effetti, per poter poi essere analizzate attraverso l'association rule mining. L'ARM può essere eseguito attraverso diversi algoritmi, quello più performante da utilizzare, in un sistema dove si richiede la frequenza dei pattern, è l'FP-Growth. L'estrazione delle regole di associazioni, dai dati forniti dal dataset della FMECA, permetterà dunque un ampliamento delle conoscenze esistenti dal sistema analizzato.

In fine si passa all'applicazione della Social Network Analysis, dove vengono posti gli attori come gli itemset frequenti, identificati dall'algoritmo e le loro interazioni come le regole, in modo da poter rappresentare graficamente, per una miglior visualizzazione, le modalità di fallimento, gli effetti e le criticità. Inoltre, sarà possibile definire sia le eventuali connessioni mancanti relative alla FMECA, sia se sono state eseguite adeguatamente le analisi dei possibili guasti.

CAPITOLO 4

IL CASO STUDIO

In questo capitolo verrà esposto il lavoro che è stato effettuato per estrarre delle regole, tramite il programma di RapidMiner. Da un dataset di partenza di tipologia FMECA, è stato possibile arrivare ad una rappresentazione grafica delle regole, grazie ai modelli del programma gephi.

4.1 DATASET DI PARTENZA

Il dataset di partenza è la rappresentazione di tutte le informazioni relative alla FMECA di un'azienda; è composto da 17 famiglie di elementi e ognuna della quale ne contiene 152.

Le famiglie sono:

1. Seq: indica la numerazione degli elementi;
2. System ID: rappresenta la numerazione del system;
3. System;
4. Subsystem ID: è la numerazione del system;
5. Subsystem;
6. Subsystem / Component ID;

7. Name:è il nome della componente analizzata;
8. MainFunction:ossia la funzione della componente del macchinario;
9. MTTR(hours): indica il rapporto tra il tempo richiesto per ripare il macchinarioe il tempo in cui rimarrà fermo, è importante tenerne conto fin da subito poiché indipendentemente dalle modalità, tutte le componenti di un macchinario sono soggette a rottura, perciò si necessiterà di riparazioni;
10. Failure Rate (λ) chosen for the equipment;
11. Odf: più grande è questo indicatore tanto più la possibilità che si presenti una situazione critica è elevata, rendendo dunque più difficile il funzionamento;
12. Mdf: se il valore di questa famiglia è molto elevato, allora si presenterà una situazione difficile per poter eseguire una manutenzione;
13. System Availability Impact: maggiore è questo indicatore, maggiore sarà l'impatto del guasto della componente sulla disponibilità dell'intero sistema;
14. Impact on People: tanto più sarà alto il valore di questo indicatore, tanto più si incorrerà in una diminuzione della sicurezza delle persone nell'eventualità in cui si presenti un guasto, durante il funzionamento e la manutenzione;
15. Environmental impact: se questo indicatore ha un valorecospicuo, le conseguenze che saranno riscontrate sono nella minor sicurezza che avranno

le persone a lavorare con quella determinata componente, durante il funzionamento e la manutenzione;

16. Component with Redundancy: indicatore binario dove i valori indicano la ridondanza:
    a. 0 se è una componente che presenta ridondanza;
    b. 1 se invece la componente non presenta ridondanza;
17. Original MPN: più è alto il valore e più e elevata la probabilità di anticipare un guasto.

4.2 RAPIDMINER

RapidMiner è un software scritto con linguaggio di programmazione java, avente come funzionalità quelle di:

- preparazione dei dati;
- machine learning;
- apprendimento profondo;
- text mining;
- analisi predittiva.

RapidMiner fornisce un'interfaccia grafica formata da flussi di lavoro analitici, che vengono chiamati "processi", i quali sono costituiti da più "operatori". Gli operatori sono di diversa tipologia ed ognuno di loro svolge un'attività differente; essi

vengono presentati con un input e un output, dove quest'ultimo, andrà a costituire l'input di quello successivo.

Questo software è uno strumento molto duttile, infatti presenta una "profondità" e complessità maggiore per chi ne usufruisce per un lavoro più specifico, come i data scientists, ma al tempo stesso, ha anche opzioni di lavoro più semplici, per chiunque sia interessato ad interagire con questa tipologia di scienza dei dati senza una profonda conoscenza della materia. Questo programma riesce dunque a trovare spazio in molteplici campi come in quello: aziendale, commerciale, di ricerca, dell'istruzione, della prototipazione rapida e dello sviluppo di applicazioni.

4.3 GEPHI

Gephi è un software open source per l'analisi dei grafici e delle reti, caratterizzato da un'architettura flessibile e multi-task, la quale garantisce di lavorare con insiemi di dati complessi e di produrre risultati visivi.

Le visualizzazioni sono state strutturante in modo da poter sfruttare le abilità percettive degli esseri umani per riuscire a trovare caratteristiche nella struttura della rete e nei dati.

Un modo per migliorare la visualizzazione, in modo da renderla più visibile e chiara e quello di applicare dei moduli grafici:questi prendono un insieme di nodi in input e modificano i parametri di visualizzazione; esempio di questi moduli sono:

- Size Gradient: serve per la grandezza del nodo, poiché più questo è grande e più potrebbe risultare importante per il sistema;
- Color Gradient: va a modificare il colore del nodo, ad esempio se la tonalità di un colore è calda, potrebbe significare che c'è la possibilità di incorrere in nodi di gravità maggiore;
- Color Clusters: modifica il colore di un gruppo di nodi, si possono indicare ad esempio, con colori diversi, dei quartieri durante lo studio di una rete sociale;
- Centrality: viene utilizzato per poter vedere quanto è buona la connessione con un nodo.

L'obiettivo è aiutare gli analisti dei dati a formulare ipotesi, scoprire in modo intuitivo schemi, isolare singolarità di struttura o di errori durante il reperimento dei dati.

Gephi, in quanto strumento di esplorazione della rete, deve garantire determinati requisiti, quali:algoritmi di layout di alta qualità, filtraggio dei dati, clustering, statistiche e annotazioni; questo deve verificarsi in modo da risultare flessibilie,

scalabile e facile da usare; queste caratteristiche consentono al software di svolgere molteplici lavori.

Date le proprietà strutturali di questo software, esso riesce a trovare campo in molti ambiti, tra i quali:

- Exploratory Data Analysis: analisi orientata all'intuizione mediante manipolazioni di reti in tempo reale;
- Link Analysis: analisi che rivela le strutture sottostanti delle associazioni tra gli oggetti;
- Social Network Analysis: analisi per la creazione di connettori di dati sociali per mappare organizzazioni comunitarie e reti di piccoli mondi.
- Biological Network Analysis: analisi per la rappresentazione di modelli di dati biologici.
- Poster Creation: promozione del lavoro scientifico con mappe stampabili di alta qualità.

4.4 APPLICAZIONE DELLE REGOLE D'ASSOCIAZIONE

Dato il dataset di partenza si è andati a definire solo un piccolo gruppo di regole di associazione, andando ad approfondire solo alcune famiglie proposte nel dataset.

La mascherina di lavoro del software RapidMiner è il foglio Process, sul quale si potranno collegare tra loro i vari operatori che il software concede; nel caso preso in esame [Figura 8] il lavoro di determinazione delle regole di associazione è lineare, quindi possiamo vedere in ordine questa serie di operatori:

1. Read Excel: questo operatore può essere usato per caricare i dati dai fogli di calcolo Microsoft Excel, sui quali era presente il dataset FMECA con tutte le famiglie;
2. Generate Attributes: questo operatore costruisce nuovi valori dando gli attributi del dataset di input e dando costanti arbitrarie: riesce a svolgere queste operazioni usando espressioni matematiche. Questo lavoro risulta necessario in quanto la famiglia "Failure Rate (λ) chosen for the equipment" non presenta lo stesso numero di elementi delle altre due che si andranno ad analizzare: sarà dunque necessario creare una nuova famiglia, che in questo caso è stata chiamata "Failure Rate (λ) chosen for the equipment 2", che verrà riempita con i valori del dataset;

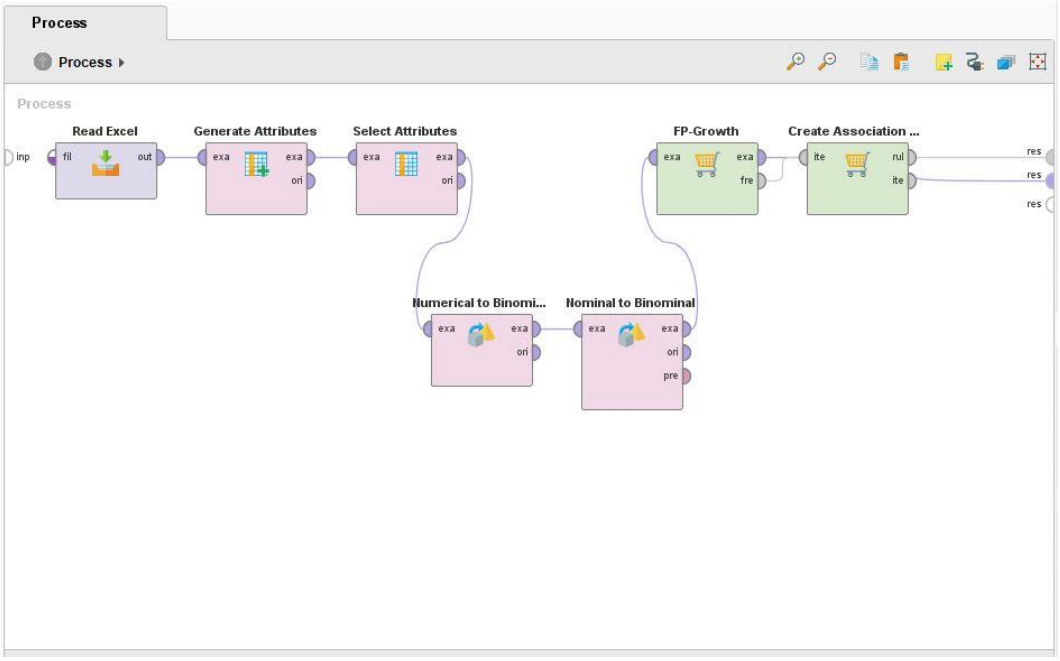

Figura 11Foglio Process del programma Rapid Miner per determinare le regole di associazione dato un dataset di partenza di tipo FMECA

3. Select Attributes: questo operatore fornisce diversi tipi di filtro per facilitare la selezione degli attributi. Grazie a lui si potranno dunque limitare le informazioni e si riuscirà a focalizzare l'attenzione solo sulle regole che ci interessano, ovvero: Failure Rate (λ) chosen for the equipment 2, Name, System;
4. Numerical to Binominal: questo operatore cambia il tipo di attributi numerici in un tipo binomiale (chiamato anche binario);

5. Nominal to Binominal: questo operatore è usato per cambiare il tipo di attributi nominali in un tipo binomiale;
6. FP-Growth: questo operatore svolge l'algoritmo FP-Growth; grazie ad esso sarà possibile determinare dunque dei pattern frequenti; inoltre con l'ausilio dei parametri che ci concede tale algoritmo sarà possibile definire il tipo di regola si cerca: i parametri presi in esame sono stati:
    a. min support;
    b. min number of itemset.
7. Create Association Rules: questo operatore crea tutte le regole di associazione presenti.

Qui, nella parte finale della struttura costruita sulla maschera di lavoro, si può aggiungere il parametro del "min support".

Una volta effettuato tutto il passaggio si potrà far partire il programma che ci mostrerà tutte le regole di associazione estratte, servirà poi spostarle in un file Excel così da poter permettere all'altro programma, GEPHY, di determinare il grafico della rete.

Premise	*Conclusion*	*Support*	*Confidence*
System = TURBINE	Failure Rate (λ) 2 = Failure Rate (λ) =0.0000228	0.039473684210526314	0.21428571428571427
Name = Water filter	System = RV	0.006578947368421052	1.0
Name = Water Temperature Sensor (cool &lub oil - RV)	System = RV	0.006578947368421052	1.0
Name = Water Temperature Sensor (cool &lub oil - RV)	Failure Rate (λ) 2 = Failure Rate (λ) =0.00001405	0.006578947368421052	1.0
Name = Distributor Valve	System = RV	0.013157894736842105	1.0
Name = Water Piping (Auxiliary Services)	Failure Rate (λ) 2 = Failure Rate (λ) =0.000000003	0.006578947368421052	1.0
Name = Upper Ring (pre-distributor)	System = TURBINE	0.006578947368421052	1.0

Name = Upper Ring (distributor)	System = TURBINE	0.0065789473684 21052	1.0
Name = Upper Guide Bearing Segment	System = AXIS	0.0065789473684 21052	1.0
Name = Upper Guide Bearing Segment	Failure Rate (λ) 2 = Failure Rate (λ) =0.0000684	0.0065789473684 21052	1.0
Name = Upper Guide Bearing Housing	System = AXIS	0.0065789473684 21052	1.0
Name = Upper Guide Bearing Housing	Failure Rate (λ) 2 = Failure Rate (λ) =0.0000684	0.0065789473684 21052	1.0
Name = Turbine Shaft	System = AXIS	0.0065789473684 21052	1.0
Failure Rate (λ) 2 = Failure Rate (λ) =0.0000684	System = AXIS	0.0460526315789 47366	0.7

Tabella 4 Parte del gruppo di regole ricavate tramite il software Rapid Miner

Nella Tabella 4 vediamo alcune delle regole ricavate da questa procedura, leggendo la tabella possiamo dedurre ad esempio che:

a) nella prima riga nel 3,95% dei casi, secondo il supporto, quando la famiglia System è di tipo TURBINE, il Failure Rate 2 avrà un valore di 0,0000228, apparendo nelle transizioni che contengono quel tipo di System un numero di volte pari al 21,43%;
b) nell'ultima riga nel 4,61% dei casi, secondo il supporto, quando la famiglia Failure Rate ha un valore di 0,0000684, il System sarà di tipo AXIS, apparendo nelle transizioni che contengono quel valore di Failure Rate un numero di volte pari al 70%. In questa maniera questa regola, tra quelle scelte ed inserite nella tabella, è quella che presenta il valore di supporto più alto e dunque la più influente; oltre a questo presenta anche un buon valore di confidenza;
c) nella seconda riga nel 0,66% dei casi, secondo il supporto, quando la famiglia Name è di tipo Water Filter, il System sarà di tipo RV, apparendo nelle transizioni che contengono quel tipo di Name un numero di volte pari al 100%. Questa regola presenta un supporto estremamente basso, e risulta essere quindi una regola estremamente poco rilevante, sebbene presenti il valore massimo di confidenza;
d) nella terza riga nel 0,66% dei casi, secondo il supporto, quando la famiglia Name è di tipo Water Temperature Sensor, il System sarà di tipo RV apparendo nelle transizioni, che contengono quel tipo di Name, un numero di volte pari al 100%. Gli indicatori della seconda e della terza riga

presentano valori praticamente identici, però non sono sintomo di una ridondanza, in quanto esaminando la riga 4 notiamo che il collegamento tra la famiglia Name e il tipo RV della famiglia System presenta valori di supporto differenti, dimostrando la non relazione con l'antenato;

e) nella quarta riga nel 1,32% dei casi, secondo il supporto, quando la famiglia Name è di tipo Distributor Valve, il Systemsarà di tipo RV, apparendo nelle transizioni, che contengono quel tipo di Name, un numero di volte pari al 100%.

Nell'esempio della terza riga si è parlato del termine ridondanza: quando si lavora con dataset estremamente grandi è facile incorrere in regole ridondanti: ciò avviene se si presentano relazioni particolari con l'"antenato", ovvero se il supporto e la confidenza della regola sono vicini al valore che possiede il suo antenato. Un esempio di questa caratteristica di alcune regole è:

{A, B, C} → {D} e {A, B} → {D} con supp=X e conf=Y per entrambe.

Un lavoro importante per diminuire il carico di regole associative è quello di riuscire a trovare queste ridondanze ed "eliminarle", dato che non saranno rilevanti ed occuperebbero solo spazio.

Una volta che si è finito il file Excel con tutte le regole, si andrà a lavorare alla rappresentazione: nel software Gephi, verranno infatti disposte in forma di grafo per poi essere riorganizzate in base alle esigenze. Inizialmente si lavorerà con lo

strumento di layout, il quale permette di dare un ordine alla struttura che si sta creando, utilizzando variabili come: repulsionstrength, maximum displacement e gravity. Si passerà poi all'utilizzo dello strumento Appearence, il quale modifica la visualizzazione di nodes e edges, in termini di colore e spessore dell'elemento oltre che di colore e grandezza delle lettere.

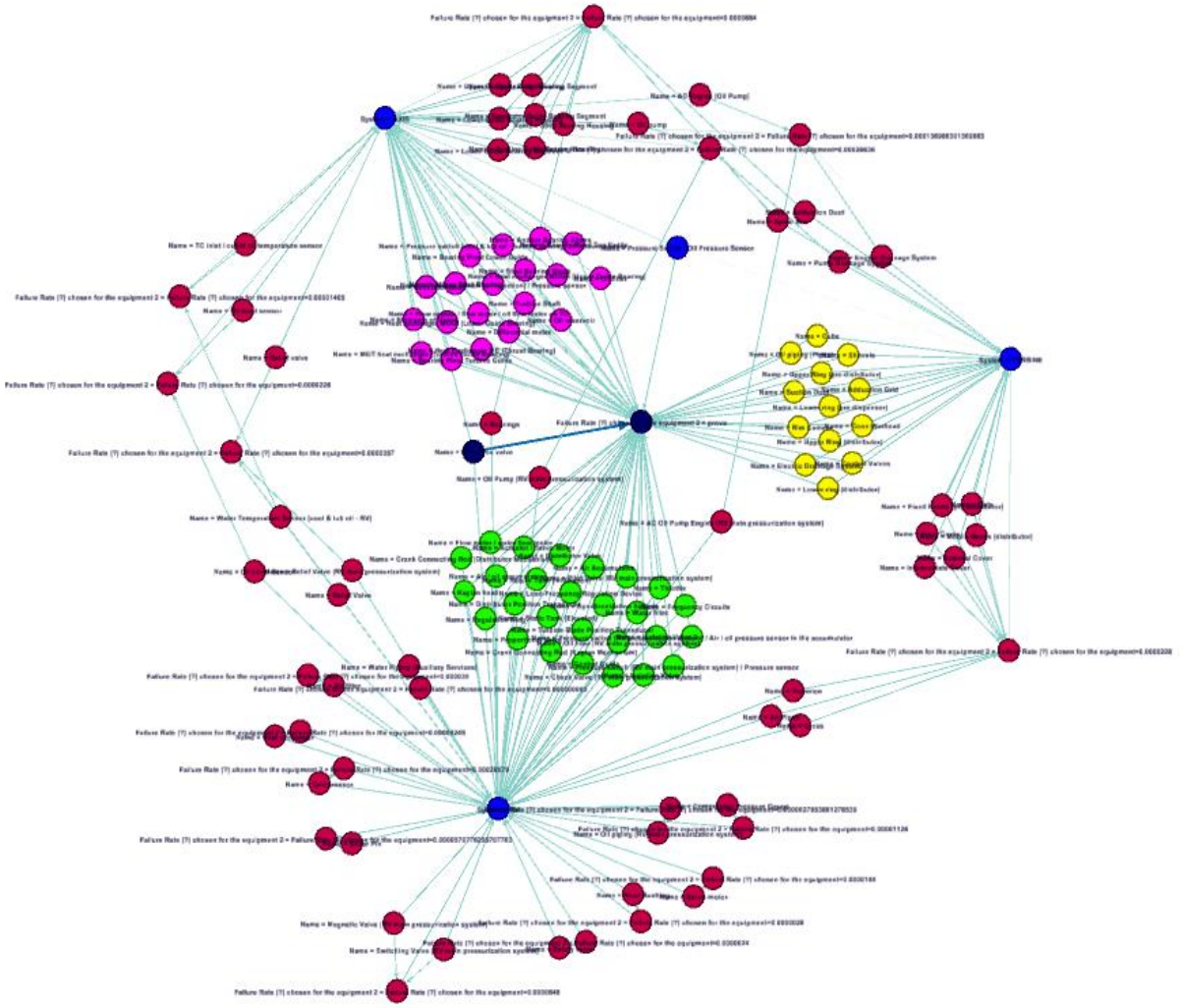

Figura 12 Grafico estratto dal software Gephi con un dataset FMECA

Nella figura 12 si vede come sono stati evidenziati tramite colori differenti (verde, giallo e viola), i 3 diversi quartieri presi in esame, si nota anche che col blu sono stati evidenziati i valori più centrali per il collegamento a vari grappoli di nodi avendo delle funzioni di ponte tra parti altrimenti inaccessibili.

CONCLUSIONE

L'utilizzo delle regole di associazione, serve per determinare collegamenti nascosti tra le componenti che vengono prese in esame: in questo caso lavorando con i dataset della FMECA, le regole prese in esame sono state funzionali a determinare le potenzialità dei gusti, i loro effetti e la criticità. Il tutto sarà reso possibile dall'utilizzo di due importanti software:

- il primo è RapidMiner, il quale presenta la parte macchina nella cooperazione con l'uomo, grazie a lui sarà dunque possibile l'estrazione delle regole necessarie all'analisi;
- il secondo invece è Gephi, il quale modellerà un grafico con le regole per sfruttare le qualità ricettive del cervello umano (parte umana nella cooperazione uomo-macchina).

Durante lo svolgimento del lavoro sono state prese in esame solo 3 delle famiglie (Failure Rate (λ) chosen for the equipment 2, Name, System), si è riuscito a determinare il collegamento tra il nome della componente e il sistema con il loro tasso di fallimento, in modo da rendere possibile la determinazione delle componenti con più probabilità di rottura o che comunque avrebbero comportato degli effetti negativi. In particolare le regole più significative sono:

1. *System = AXIS → Failure Rate (λ) chosen for the equipment 2 = 0,0000228310502283105*

presentando un supporto del 26,3% e una confidenza del 45,5%.

2. *Failure Rate (λ) chosen for the equipment 2 = 0,00002283105022831 05 → System = AXIS*

   si presenta come la regola inversa alla precedente (presentando dunque degli archi a doppia freccia durante una rappresentazione grafica), ed è formata dallo stesso supporto con una confidenza minore (57,1);

3. *System = RV → Failure Rate (λ) chosen for the equipment 2 = 0,00014269406392 6941*

   presentando un supporto del 21,1% e una confidenza del 59,3%.;

Analizzando la numero 3 vediamo come il valore del *Failure Rate (λ)* si presenta di un'unità decimale superiore alle regole che lo precedono e dato che mantiene dei valori di supporto e confidenza leggermente simili (minore supporto e maggiore confidenza), possiamo definire questa regola come la più importante per lo specifico studio fatto.

BIBLIOGRAFIA

- http://tesi.luiss.it/17593/1/078262_FUGGITTI_FILIPPO.pdf
- https://www.mdpi.com/1996-1073/13/23/6400
- https://dl.acm.org/doi/10.1145/3230485
- http://www.columbia.edu/~rd2537/docu/apriori(abstract).pdf
- http://bias.csr.unibo.it/golfarelli/DataMining/MaterialeDidattico/DMISI-RegoleAssociative.pdf
- https://www.politesi.polimi.it/handle/10589/121041
- https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/
- https://gephi.org/
- http://rakesh.agrawal-family.com/papers/vldb95tax_rj.pdf
- https://onlinelibrary.wiley.com/doi/pdf/10.1002/1099-1638%28200007/08%2916%3A4%3C313%3A%3AAID-QRE434%3E3.0.CO%3B2-U
- https://doi.org/10.1007/BF02345483