ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

FACOLTA' DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

DIPARTIMENTO DI INFORMATICA SCIENZA E INGEGNERIA

TESI DI LAUREA

in

Processi e tecniche di Data Mining M

Analisi di dati citofluorimetrici con tecniche di Data Mining

CANDIDATO
Riccardo Cova

RELATORE:
Chiar.mo Prof. Ing. Claudio Sartori

CORRELATORI
Ing. Davide Sottara

Anno Accademico 2011/12

Sessione II

Parole chiave: ***Data Mining***
Clustering
Classificazione
K-Means
Expectation Maximisation
DBscan
SVM
SMO
Java
Weka
Eclipse
Weasel
Biologia
Citofluorimetria a flusso
Citofluorimetro
Spermatozoo

Sommario

Introduzione

Il citofluorimetro è uno strumento impiegato in biologia genetica per analizzare dei campioni cellulari: esso, analizza individualmente le cellule contenute in un campione ed estrae, per ciascuna cellula, una serie di proprietà fisiche, *feature*, che la descrivono.

L'obiettivo di questo lavoro è mettere a punto una metodologia integrata che utilizzi tali informazioni modellando, automatizzando ed estendendo alcune procedure che vengono eseguite oggi manualmente dagli esperti del dominio nell'analisi di alcuni parametri dell'eiaculato.

Si cerca di migliorare l'analisi dell'eiaculato in citofluorimetria per ottenere i seguenti vantaggi:

- risultati più accurati
- risultati più oggettivi
- riduzione dei costi
- possibilità di eseguire le analisi in qualunque luogo in cui sia presente un citofluorimetro a flusso tradizionale
- drastica riduzione dei tempi di analisi
- analisi di un elevato numero di spermatozoi del campione

Questo richiede pertanto lo sviluppo di tecniche biochimiche per la marcatura delle cellule e tecniche informatiche per analizzare il dato.

Il primo passo prevede la realizzazione di un classificatore che, sulla base delle *feature* delle cellule, classifichi e quindi consenta di isolare le cellule di interesse per un particolare esame.

Il secondo prevede l'analisi delle cellule di interesse, estraendo delle *feature* aggregate che possono essere indicatrici di certe patologie. Il requisito è la generazione di un report

esplicativo che illustri, nella maniera più opportuna, le conclusioni raggiunte e che possa fungere da sistema di supporto alle decisioni del medico/biologo.

Cap. 1 - Citofluorimetria a flusso

La citofluorimetria a flusso è largamente utilizzata nella ricerca biomedica come nei lavori svolti sulle cellule staminali o nello sviluppo dei vaccini. Clinicamente è impiegata per monitorare il corso e il trattamento delle infezioni da HIV e nella diagnosi e monitoraggio delle leucemie e linfomi *[1]*.
Per anni la citofluorimetria a flusso è stata utilizzata in pochi laboratori, ma il continuo miglioramento delle performance dei computer e la riduzione del loro prezzo di acquisto hanno permesso di utilizzare efficacemente questa tecnica, poiché attualmente è possibile analizzare centinaia di campioni ogni giorno.

Tradizionalmente l'analisi dei dati forniti dalla citofluorimetria a flusso richiedeva grande dispendio di tempo, in quanto veniva eseguita manualmente su grafici multicolor *[15]*.
Nei testi clinici questo era causa di variazione nei risultati, infatti l'esperto del dominio nel fare l'analisi si basa sulla sua esperienza e sulla sua intuizione piuttosto che sull'applicazione delle tecniche di inferenza statistica sui dati.
L'evoluzione della citofluorimetria a flusso ha messo a disposizione degli esperti sempre più dati da poter analizzare e questo ha richiesto lo sviluppo di tecniche statistiche affidabili e relativi software di gestione, che potessero essere di supporto per eseguire un'analisi più accurata, riproducibile e standard [1].

1.1 Il citofluorimetro

Il citofluorimetro è uno strumento simile a un microscopio, nel quale però si lavora su oggetti spinti da un flusso attraverso un capillare verso la cella di flusso che rappresenta il punto in cui si focalizza il raggio laser a 488nm che illumina le cellule. Queste passano una in fila all'altra davanti al laser che come prima cosa ne ricava due parametri fisici: Forward Scatter (FS) e Side Scatter (SS). Questi due parametri misurano come l'oggetto devia la luce del raggio laser con il quale viene illuminato.
Il rilevatore per il FS è un fotodiodo posto a circa 180° rispetto alla direzione del raggio laser e misura la quantità di fotoni lasciati passare dall'oggetto proporzionalmente alla sua dimensione e forma.

Il rilevatore per il SS è invece un fotomoltiplicatore con un diverso livello di sensibilità rispetto al fotodiodo, e misura la quantità di fotoni che vengono deviati lateralmente in funzione della densità e granularità della cellula; il fotomoltiplicatore si trova circa a 90° rispetto al raggio laser.
FS e SS combinati insieme danno quindi un'informazione di tipo fisico: misurano la dimensione dell'oggetto e la sua complessità.

Per poter avere un misura precisa delle dimensioni degli oggetti solitamente si creano delle sferette di dimensione nota e le si fanno passare attraverso il citofluorimetro misurandone il valore di FS e SS. Con questi valori verranno poi confrontati FS e SS delle cellule del campione e se ne dedurranno, con una proporzione, le dimensioni reali.

Il citofluorimetro è provvisto di ulteriori fotomoltiplicatori adibiti alla misura della fluorescenza emessa dalle cellule colorate con una molecola fluorescente. Una molecola fluorescente può assorbire un fotone di luce di una certa lunghezza d'onda e riemettere un fotone a una lunghezza d'onda più lunga (ad esempio una molecola illuminata con una luce verde emetterà il colore rosso).
Nel citofluorimetro questi fotomoltiplicatori[1] sono indicati con FLn:

- FL1 misura la luce verde
- FL2 misura la luce giallo arancio
- FL3 misura la luce rossa
- FL4 misura la luce rossa scura vicino all'infrarosso

Questi fotomoltiplicatori sono degli oggetti regolabili, cioè è possibile cambiarne la sensibilità aumentando o diminuendo il voltaggio o il guadagno.
Come per le misure di SS e FS anche per la fluorescenza solitamente si tara il citofluorimetro preparando delle sferette con una quantità di fluorescenza nota e una volta analizzate dal citofluorimetro ottenere una scala di fluorescenza con la quale confrontare le cellule dei campioni analizzati.

[1] Spesso viene misurata la luce verde poiché molti dei fluorocromi che emettono la luce verde sono eccitati molto bene dal raggio laser a 488nm (che produce una luce blu) utilizzato nel citofluorimetro.

1.2 Cellule eucariote

Una cellula eucariota, caratteristica dei mammiferi compreso l'uomo, è composta da una membrana esterna plasmatica e da un nucleo che contiene il materiale genetico (DNA e RNA). Il nucleo è a sua volta rivestito dalla membrana nucleare. Tra la membrana nucleare e quella esterna si trova il citoplasma composto da svariati organuli come i mitocondri e i ribosomi. Il materiale all'interno della membrana nucleare è solitamente molto più denso del citoplasma.

1.2.1 Il DNA nelle cellule

La quantità di DNA presente in una cellula viene misurata in un'unità di misura arbitraria. Nell'uomo la quantità di DNA cellulare è pari a 2C (7 picogrammi) poiché ogni cellula sana possiede 23 coppie di cromosomi. Con il valore di 1C (pari a 3.5 picogrammi) viene invece indicata la quantità di DNA presente normalmente nello spermatozoo umano; questo valore deriva dal fatto che nello spermatozoo la quantità di cromosomi, rispetto alle altre cellule, è la metà; solo fondendosi con l'ovocita (che fornisce i cromosomi materni) si ottengono i 46 cromosomi e quindi la quantità 2C.
Gli spermatozoi vengono chiamati cellule aploidi mentre tutti gli altri tipi di cellule vengono chiamate cellule diploidi.

1.2.2 La misura del DNA nelle cellule

In ambito cellulare, particolarmente interessante è la misurazione della quantità di DNA presente a livello nucleare, aspetto assai importante da diversi punti di vista: cellule non tumorali presentano la stessa quantità di DNA, mentre solitamente una cellula tumorale aumenta o diminuisce il suo normale contenuto di DNA. Attraverso il citofluorimetro è possibile analizzare il sangue e ricercare ad esempio un gruppo particolare di linfociti; se una certa percentuale di linfociti presenta una determinata percentuale di DNA oltre il valore normale è possibile dedurre la presenza di una leucemia.

Per misurare la quantità di DNA si usano delle molecole fluorescenti che si legano alla doppia elica del DNA e la colorano in modo specifico.

Una caratteristica desiderabile di una molecola fluorescente è quella di riuscire a legarsi in modo stechiometrico, cioè lineare alla quantità di DNA che è presente nella cellula. Questo permette di dedurre la quantità di DNA presente nella cellula, misurando la quantità di fluorescenza emessa dalla cellula quando viene illuminata.

Per colorare il DNA una sostanza molto utilizzata è il Propidio Ioduro per via della sua particolare caratteristica che deriva dal modo in cui si lega alla doppia elica del DNA. Il DNA è infatti composto da uno scheletro esterno di zuccheri e fosfati; all'interno di questa struttura ci sono delle basi che si legano con legami specifici; tra una base e quella successiva vi è dello spazio in cui può inserirsi una sola molecola di Propidio Ioduro. Una volta riempiti tutti gli spazi nessuna ulteriore molecola di Propidio Ioduro può rimanere legata alla cellula. Successivamente misurandone la fluorescenza, una volta tarato il citofluorimetro, è possibile risalire alla quantità di DNA presente all'interno della cellula.

1.2.3 La misura del danno al DNA della cellula

Affinché lo spermatozoo possa impiantarsi nell'ovulo e successivamente non portare ad un aborto, il suo DNA deve essere integro. Per verificare[2] ciò si misurano i danni fisici allo scheletro del DNA, come l'interruzione della doppia elica.
I danni ipotizzabili sono due:

- SSB (Single Strand Break): un'interruzione su un singolo filamento (l'ulteriore filamento rimane intatto)
- DSB (Double Strand Break): interruzione in entrambi i filamenti.

Per entrambi i problemi la metodica di individuazione è la stessa: si marcano i terminali dei segmenti rotti dell'elica del DNA con una molecola fluorescente, l'Isotiocianato di Fluoresceina (FITC), attraverso una reazione chimica. Al crescere dei tratti di DNA danneggiati maggiore sarà la quantità della molecola a legarsi ad esso e maggiore sarà la fluorescenza emessa dalla cellula indicando la quantità di danno al DNA.

[2] Si potrebbe pensare di misurare la quantità di DNA nello spermatozoo e verificare se è diversa dalla quantità normale 1C ed in tal caso presumere che lo spermatozoo non sia sano, ma questa procedura non è seguita dagli esperti perché risulta molto complessa.

La fluorescenza di questa molecola, diversa da quella utilizzata per marcare il DNA, viene successivamente letta da un apposito fotomoltiplicatore FLn.

1.2.4 Marcatura specifiche e aspecifiche

La marcatura con la molecola fluorescente avviene attraverso una reazione biochimica specifica nella quale la molecola si lega al filamento di DNA nel punto in cui è presente il danno. In teoria si dovrebbero marcare esclusivamente i punti del filamento danneggiati, nella pratica una piccola percentuale[3] di questa molecola si lega a tutti[4] gli spermatozoi, anche nelle posizioni in cui non è presente il danno. Questo è definibile come rumore e viene chiamato legame aspecifico.

Per individuare il corretto valore di danno al DNA le analisi vengono svolte secondo questa procedura:

- il campione viene diviso a metà
- in una metà viene aggiunta la molecola fluorescente ma non viene attivato l'enzima che scatena la reazione biochimica; in questo campione (campione di controllo) il colorante si sarà legato alle cellule, non a causa del danno al DNA, ma perché vi è rimasto "intrappolato".
- nell'altra metà del campione (campione di test) si fa avvenire la reazione chimica; in questo caso verranno marcati tutti i danni alla struttura del DNA della cellula.
- per ottener il dato finale si sottrae *[4]* il campione di controllo dal campione di test; in questo modo si esclude dal campione di test la colorazione aspecifica.

[3] Gli studi condotti dal Prof. Bizzaro sono andati nella direzione di individuare una molecola che permettesse una marcatura il più specifica possibile.

[4] Non esiste uno spermatozoo umano assolutamente senza danni al DNA. È l'ovocita che è in grado di riparare i danni dello spermatozoo. Il problema dell'infertilità si presenta se lo spermatozoo ha più danni di quelli che l'ovocita può riparare (e ad oggi non si è ancora scoperta la quantità di danni riparabili).

1.3 Utilizzi del citofluorimetro

Il citofluorimetro a flusso solitamente si utilizza nelle analisi del sangue ad esempio per la conta dei globuli bianchi, rossi ed altri tipi di cellule. Questo strumento permette di distinguere le varie tipologie di cellule sulla base delle loro caratteristiche di FS e SS.

Ad esempio un linfocita presenta un nucleo abbastanza compatto e nel momento in cui si presenta un'infezione si scatena una massiva produzione anticorpi. Esistono anche altri tipi di cellule tra cui i leucociti polimorfi nucleati che, invece di avere un normale nucleo, presentano un nucleo lobato. Queste due cellule a parità di dimensioni quando passano per il citofluorimetro danno dei valori di FS e SS diversi: è probabile che il valore di SS sarà simile perché hanno la stessa dimensione, ma il valore di FS sarà diverso poiché questo parametro è influenzato dalla complessità del nucleo interno. Ciò permette di distinguere i due diversi tipi cellulari.

1.4 Interessi verso lo studio della fertilità maschile

Lo studio approfondito degli spermatozoi di una popolazione di individui, oltre ai fini prettamente riproduttivi, è collegato allo stato di "salute" dell'ambiente in cui la popolazione vive. Infatti le alterazioni degli spermatozoi sono un buon indice dello stato di salute generale del paziente poiché la loro produzione è un processo molto sensibile e complicato, particolarmente suscettibile a moltissime alterazioni dello stato di salute fisico ma anche all'interazione con gli agenti inquinanti presenti nell'ambiente circostante. L'analisi dello spermiogramma di una popolazione potrebbe quindi essere un sistema per rendersi conto della presenza di qualche agente inquinante e dannoso cui è esposta la popolazione. A oggi questo tipo di analisi viene svolta da parte di un esperto e richiede molto tempo; anche in questi studi si cerca di applicare tecniche di analisi statistica per ottenere i vantaggi discussi precedentemente.

1.4.1 Citofluorimetria a flusso e malattie della riproduzione

Oggi lo spermiogramma, cioè l'analisi del liquido seminale, è eseguito da parte di un esperto che al microscopio identifica visivamente e manualmente un certo numero di spermatozoi e ne valuta le loro proprietà. Questo è un processo che richiede molto tempo e costi elevati.

1.4.1.1 Lo spermatozoo

Lo spermatozoo maturo è una cellula super specializzata per la funzione riproduttiva. Per fare ciò deve trasportare il materiale genetico all'interno dell'ovocita femminile. Per poter portare l'informazione genetica lo spermatozoo deve essere il più possibile idrodinamico. Rispetto ad una cellula "classica" lo spermatozoo è privo del citoplasma e mantiene solo l'involucro esterno, la membrana plasmatica, la quale si stringe intorno alla membrana del nucleo risultando così molto compresso. Per muoversi lo spermatozoo è datato di un flagello e l'energia è prodotta dai mitocondri. Per poter introdursi nell'ovocita lo spermatozoo è dotato inoltre di enzimi che bucano la membrana dell'ovocita.

1.4.1.2 Infertilità

I problemi di infertilità possono essere diversi. Un individuo è detto azoospermico se non produce spermatozoi, oligospermico se ne produce pochi, oppure è chiamato astenospermico nel caso in cui presenti il normale numero di spermatozoi di una persona sana, ma essi sono poco mobili o presentano delle morfologie sbagliate (flagello corto, due flagelli) o alterazioni della testa.

1.4.2 Difficoltà nella analisi dello spermiogramma con citofluorimetro

1.4.2.1 Passaggio multiplo

Gli spermatozoi si trovano nell'eiaculato, ma nel campione sono presenti anche altri tipi di cellule come ad esempio i linfociti nel caso di una infezione in atto.

Queste cellule che non sono spermatozoi hanno una quantità di DNA doppia rispetto agli spermatozoi. Non è però immediato riuscire a distinguere queste cellule dagli spermatozoi poiché non è possibile garantire che nella cella di flusso del citofluorimetro passi una sola cellula per volta; la conseguenza è che una quantità di DNA pari a 2C potrebbe essere

dovuta al passaggio di una comune cellula oppure dal passaggio di due spermatozoi contemporaneamente.

Esistono comunque tecniche pre-analitiche che riducono notevolmente le probabilità di un passaggio contemporaneo di più cellule nella cella di flusso. E' anche possibile individuare [5] questo passaggio contemporaneo di cellule monitorando i valori di FF e SS che saranno per forza differenti dal passaggio di una singola cellula.

1.4.2.2 Forma degli spermatozoi

Il citofluorimetro nasce come strumento per analizzare dei campioni di forma tendenzialmente sferica come le cellule del sangue. Gli spermatozoi non avendo una forma sferica, a seconda di come attraversano la cella di flusso, saranno colpiti dal raggio laser in maniera differente producendo valori diversi di FS e SS.

Presso pochissimi laboratori specializzati sono presenti costosissimi citofluorimetri specifici[6] per l'analisi degli spermatozoi che permettono di orientare tutte le cellule nella stessa direzione durante il passaggio nella cella di flusso. Proprio a causa della presenza limitata di questo particolare tipo di citofluorimetro, ma dell'abbondanza di citofluorimetri tradizionali uno degli obiettivi di questo lavoro è permettere l'uso di tale strumento per l'analisi degli spermatozoi.

1.4.2.3 Diversità

Nei mammiferi sani gli spermatozoi sono fisicamente uno uguale all'altro e questa caratteristica viene detta spermatogenesi perfetta. L'essere umano invece, indifferentemente dal suo stato di salute, è caratterizzato da una spermatogenesi imperfetta nella quale ogni spermatozoo è diverso dall'altro. Questo fattore non determina comunque l'infertilità di un individuo.

[5] Un possibile sviluppo futuro potrebbe essere quello di indentificare e distinguere le varie tipologie di cellule che non sono spermatozoi (possono essere cellule linfocitarie o leucocitarie) perché nello studio delle malattie che minano la riproduzione è importante riuscire a riconoscerle.

[6] Questi particolari citofluorimetri vengono utilizzati per fare la selezione degli spermatozoi con un preciso cromosoma sessuale. Sono utilizzati per eseguire fertilizzazioni specifiche per ottenere solo maschi o solo femmine.

1.4.2.4 Il DNA è DNA

Le molecole fluorescenti utilizzate per marcare il materiale genetico ed evidenziarne i danni si legano al DNA delle cellule presenti nel campione indifferentemente dal tipo di cellula. Non è quindi possibile riuscire a distinguere il DNA di uno spermatozoo da quello delle altre cellule presenti nell'eiaculato.

La soluzione a questi problemi per individuare gli spermatozoi viene dall'utilizzo combinato di più parametri.

1.5 Citofluorimetria - Analisi manuale dello spermiogramma

Le metodologie di individuazione degli spermatozoi sono diverse. Nel seguito è illustrato, in maniera non esaustiva, il procedimento utilizzato dal Prof. Davide Bizzaro dell'istituto di Biologia e Genetica dell'Università Politecnica delle Marche.

Un eiaculato medio contiene 2 - 4ml di liquido nel quale sono presenti dai 20 ai 100 milioni di spermatozoi per millilitro. Nelle analisi al citofluorimetro, come in quelle svolte manualmente al microscopio, si analizza un campione dell'eiaculato, estratto dopo averlo miscelato, che comprende mediamente 40000 cellule di cui solitamente solo 10000 sono spermatozoi.

Primariamente gli spermatozoi vengono identificati attraverso la loro forma. Per eseguire una prima selezione di cellule viene riportato il risultato del citofluorimetro in grafico bi-parametrico FS-SS come illustrato in Figura 1 e Figura 2.

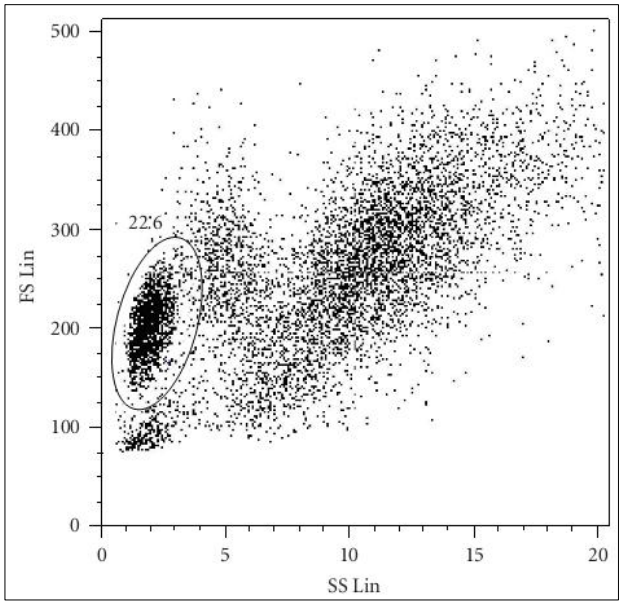


Figura 1: Grafico SS-FS delle cellule di un campione analizzato con il citofluorimetro

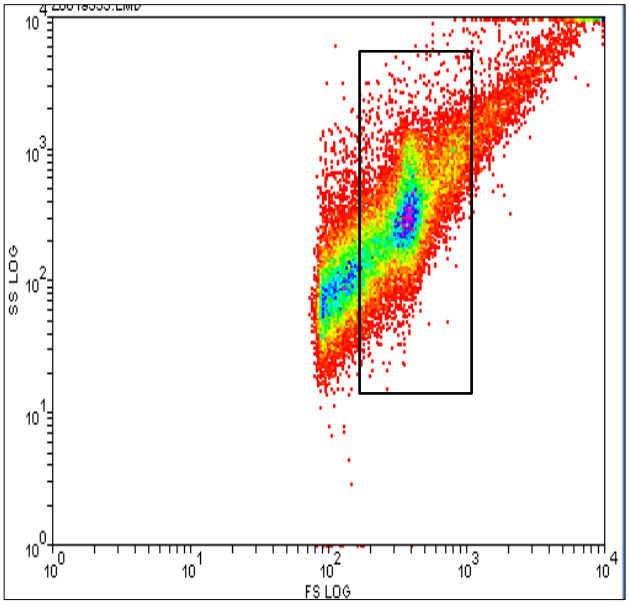


Figura 2: Grafico SS-FS delle cellule del campione di un eiaculato. Colori in base alla densità dei punti

Nel grafico è possibile individuare diverse popolazioni di cellule e classificarle in base ai valori di SS e FS. Ipotizzando che esso rappresenti l'eiaculato di un individuo che ha un'importante infezione, al suo interno possiamo trovare non solo gli spermatozoi, ma anche batteri, linfociti, macrofagi, cellule della mucosa e cellule morte.

Solitamente[7] gli spermatozoi rappresentano un gruppo di cellule molto addensato disposto al centro della nuvola (vedi Figura 2). La prima selezione (gate) che si esegue si basa quindi sui valori di SS e FS corrispondenti ad un insieme di cellule che si dispongono ad ellisse

[7] Per fini non diagnostici, ma sono allo scopo di capire come sono caratterizzati normalmente gli spermatozoi, in modo da riuscire approssimativamente ad individuarli nei grafici, si prende un campione di un eiaculato e lo si tratta con metodi fisici come la centrifugazione o il gradiente di densità ottenendo un campione contenente quasi solo spermatozoi. Una tecnica utilizzata è quella che prevede di sfruttare la caratteristica della mobilità degli spermatozoi : l'eiaculato viene messo in una provetta insieme ad un liquido di coltura; le cellule che non si muovono rimangono in fondo alla provetta, mentre gli spermatozoi tenderanno a muoversi in tutte le direzioni ed in generale a salire (swim up) nel liquido di coltura. Si preleva successivamente solo il liquido di coltura riuscendo ad ottenere un liquido con all'interno dall'80 al 99% di cellule che sono spermatozoi.
Analizzando questo campione con il citofluorimetro e visualizzandone il risultato in un grafico bi parametrico FS – SS si riscontra che le cellule, in pratica gli spermatozoi, si dispongono in una regione di spazio molto concentrata a forma di ellisse con l'asse maggiore lungo la dimensione SS ed asse minore lungo la dimensione FS.

orientato lungo l'asse SS. L'esperto traccia quindi una regione rettangolare piuttosto abbondante (per essere sicuro di includere gran parte degli spermatozoi) intorno alla forma ellissoidale come è visibile nella Figura 2.
Utilizzare però esclusivamente i valori FS e SS non è sufficiente per individuare con esattezza gli spermatozoi perché oggetti del campione che hanno valori di SS e FS compatibili con quelli degli spermatozoi non significa che siano spermatozoi.

Nella fase successiva le cellule del campione selezionate vengono allora valutate sotto un altro punto di vista: l'esperto analizza il grafico rappresentante i valori della molecola fluorescente che si lega al DNA , il Propidio Ioduro. Solitamente in questo grafico si riescono a distinguere 3 zone (vedi Figura 3):

- un primo picco (Picco 1) che identifica un ammasso di elementi molto poco o per niente fluorescenti. Gli elementi che non presentano fluorescenza, cioè che non hanno DNA, non sono cellule.
- un picco sulla destra dell'istogramma (Picco 3) che identifica un gruppo di cellule ad alto valore di fluorescenza. Queste possono essere cellule molto grandi oppure ammassi di nuclei cellulari.
- un picco centrale (Picco 2) che identifica un gruppo di cellule che si è colorato poichè contiene una certa quantità di DNA.

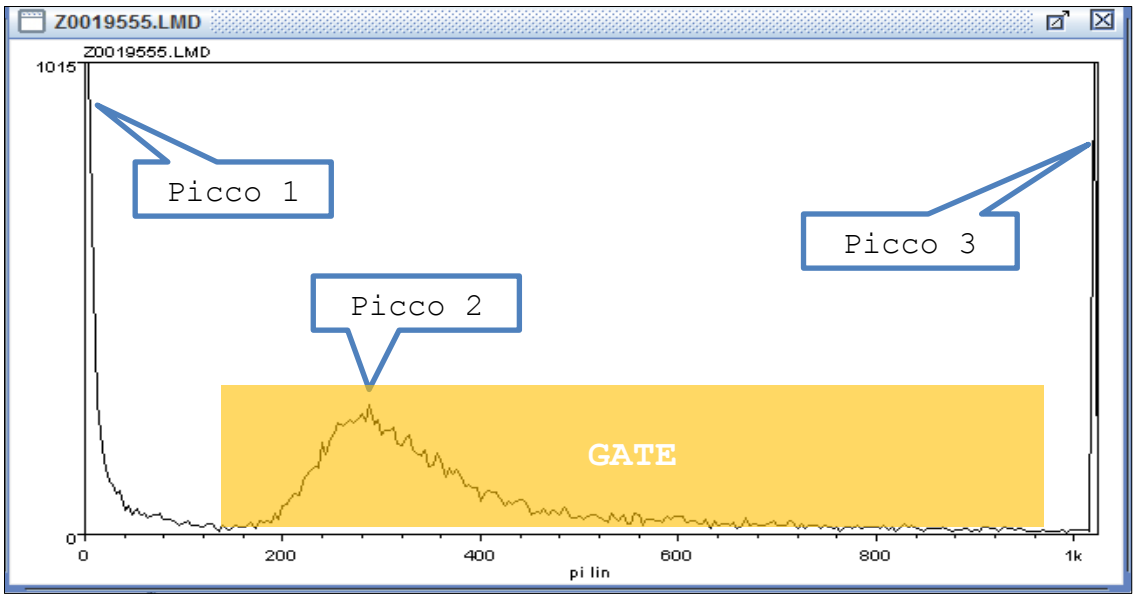


Figura 3: Grafico dei valori di fluorescenza dell'intero campione

Poiché ogni spermatozoo deve contenere una certa quantità di DNA vicino al valore 1C, analizzando la quantità di fluorescenza del Propidio Ioduro delle cellule che presentano valori compatibili di FS e SS con quelli degli spermatozoi, è possibile individuare con una buona precisione gran parte degli spermatozoi del campione.

Questa affermazione necessita però di una precisazione importante per capire la natura dei dati che vengono analizzati. Nelle cellule diploidi il DNA non è libero all'interno della membrana nucleare, ma è legato a delle proteine, gli istoni. Questo perché se il DNA fosse libero occuperebbe molto spazio e la cellula sarebbe molto grande. Per occupare il minor spazio possibile i filamenti di DNA che hanno una carica negativa sono attratti e raggomitolati sugli istoni che possiedono una carica positiva che li tiene compressi.

Nello spermatozoo questa condensazione del DNA è ancora più forte poiché deve essere efficiente e idrodinamico. Negli spermatozoi allora gli istoni sono sostituiti dalle protamine che sono più piccole ma con un potere attrattivo superiore.

Nello spermatozoo quindi il DNA è molto compattato e questo è un fattore positivo per la funzione che deve compiere, ma crea dei problemi quando si cerca di individuare la quantità di DNA al suo interno. Il Propidio Ioduro colora infatti molto bene il DNA quando questo non è molto compattato[8], ma è meno efficace quando questo è molto compresso.

A causa di ciò non è possibile individuare il valore di fluorescenza di una normale cellula somatica che contiene una quantità di DNA pari a 2C e dedurre che gli spermatozoi avranno un valore di fluorescenza pari a circa la metà perché sono cellule aploidi 1C.

Nello spermatozoo, infatti, a causa del forte compattamento del DNA che provoca un difficile accesso da parte del colorante, il valore di Propidio Ioduro che colora il DNA è inferiore alla metà di quello che colora le cellule diploidi. Questa caratteristica viene detta ipocolorazione e lo spermatozoo sembra una cellula ipoaploide, cioè una cellula che ha meno DNA di una cellula aploide.

Questo comportamento indesiderato del colorante crea delle difficoltà nella individuazione corretta degli spermatozoi perché tra le cellule che sembrano ipoaploidi vi sono anche

[8] I linfociti sono cellule che possiedono un DNA non molto condensato, quindi si colorano circa tutti allo stesso modo e nell'istogramma del Propidio Ioduro si identificano da un picco abbastanza netto

frammenti di cellule o cellule che proprio per loro natura sono ipoaploidi perché ad esempio hanno perso dei cromosomi[9].
Inoltre a causa della spermatogenesi imperfetta, la quale comporta che ogni spermatozoo sia diverso dagli altri e quindi sia diverso anche il livello di condensazione[10], la quantità di colorante che si lega al DNA è variabile in ogni spermatozoo.

A causa di questa variabilità di colorazione si esegue un gate con un intervallo di valori di fluorescenza molto ampio.
Solitamente la selezione della finestra di intervalli del valore di fluorescenza avviene individuando come valore minimo quello corrispondente al punto di minimo nel grafico tra il "Picco 1" e il "Picco 2", mentre il valore massimo di fluorescenza è scelto tra il canale (valore di fluorescenza) 800 ed il canale 1000, in ogni caso escludendo l'ultimo picco.

A conferma di questa tecnica di individuazione manuale degli spermatozoi nelle figure seguenti è illustrato il gate eseguito sull'intero campione valutando il Propidio Ioduro (vedi Figura 4) e poi la visualizzazione delle sole cellule del gate stesso nel dot plot FS – SS (vedi Figura 6).

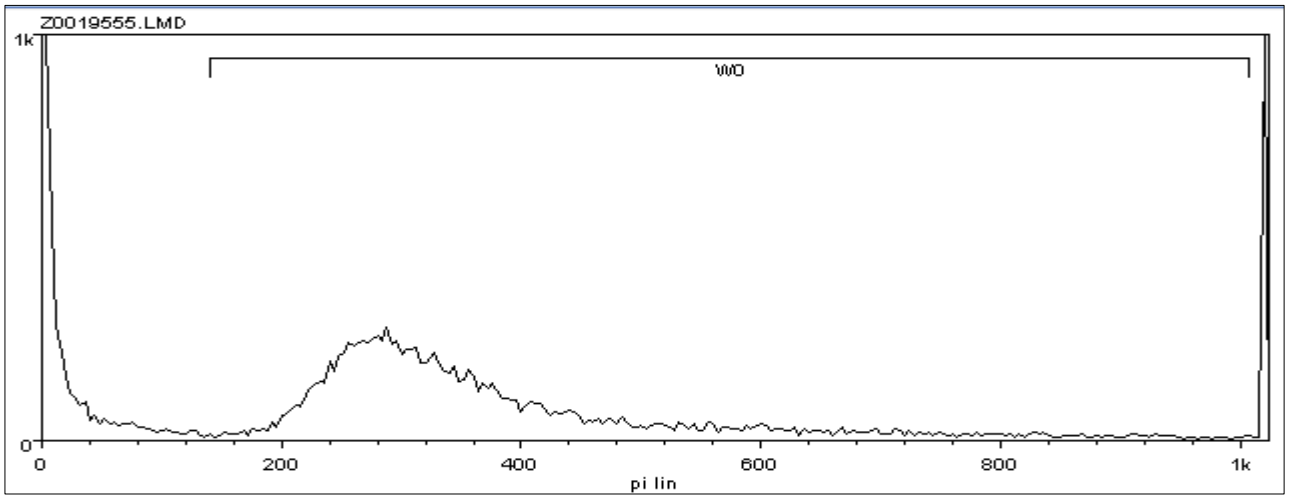


Figura 4: Selezione della regione di interesse (gate) nel grafico del Propidio Ioduro delle cellule dell'intero campione

[9] In teoria come caso limite è possibile che uno spermatozoo danneggiato che ha perso alcuni cromosomi si colori tutto, perché è meno condensato, e quindi risulti più fluorescente di uno spermatozoo sano con tutti i cromosomi che però si è colorato meno a causa del fatto che ha il suo DNA molto più condensato.
[10] La condensazione è un processo fondamentale per la funzionalità degli spermatozoi, ma una alta presenza di spermatozoi ipocondensati o ipercondensati è percentualmente più presente in individui con problemi di infertilità. L'accessibilità del colorante è quindi un parametro di valutazione della qualità degli spermatozoi di quell'individuo.

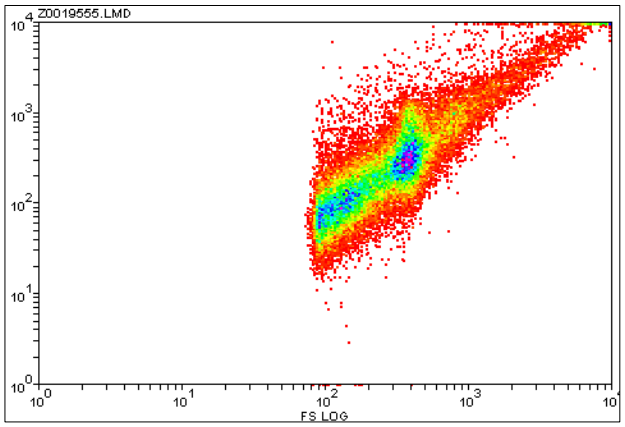


Figura 5: Intero campione

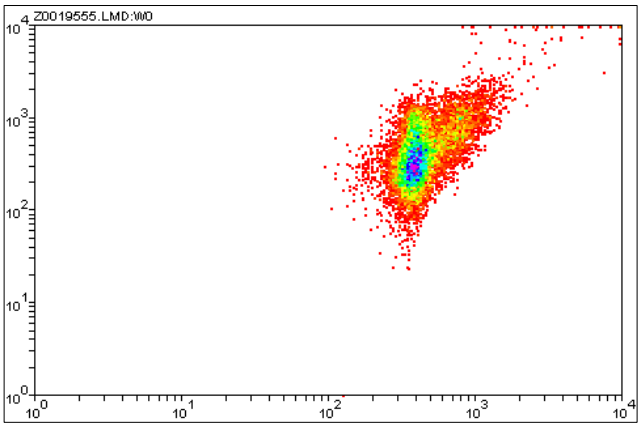


Figura 6: Gate visualizzato nel dot plot FSLog-SSLog

Le cellule appartenenti al gate sono come supposto appartenenti alla nuvola centrale a forma di ellisse. Il grafico visualizza anche un certo numero di cellule piuttosto grandi (valori di FS e SS elevati) che potrebbero essere linfociti, spermatozoi con all'interno il doppio della quantità di DNA o cellule che passano insieme nel citofluorimetro.

Dal gate eseguito sul grafico SS-FS seguito dal gate sul Propidio Ioduro si ottiene un insieme di cellule che per l'esperto rappresentano gli spermatozoi. Il passo finale dell'analisi è quello di individuare e valutare il danno al DNA andando a graficare la misura della molecola fluorescente che si lega alle strutture interrotte dei filamenti di DNA. Il grafico illustrato in Figura 7 è un grafico di distribuzione di frequenze che mostra sulle ascisse la quantità di fluorescenza relativa alla molecola che si lega alle strutture danneggiate (FITC) e sulle ordinate il numero di cellule.

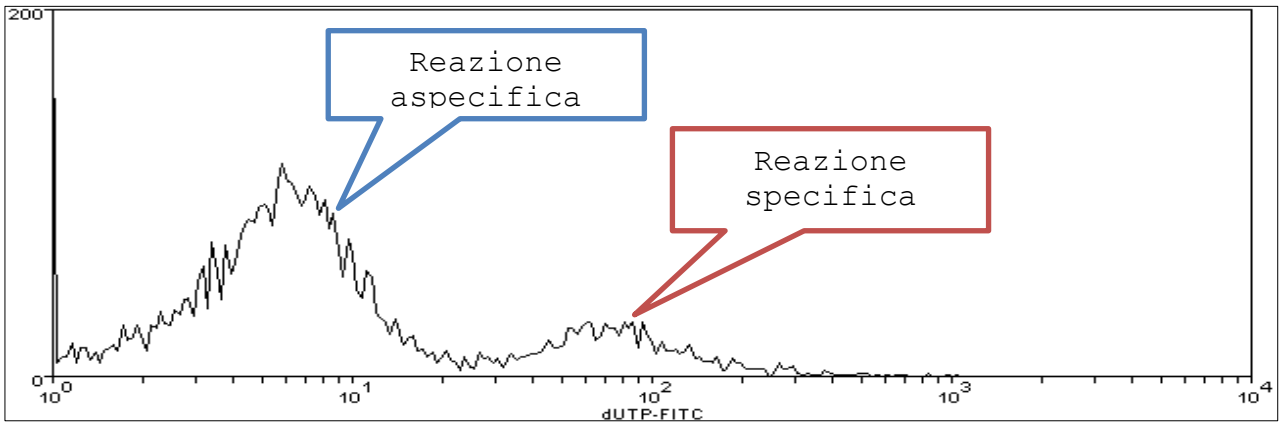


Figura 7: Danno al DNA degli spermatozoi

Tipicamente un individuo con DNA danneggiato negli spermatozoi presenta una distribuzione simile a quella illustrata in Figura 7 nella quale è evidente un picco marcato di bassa fluorescenza dovuto tendenzialmente al colorante che si lega in maniera aspecifica, e un altro picco con valori di fluorescenza più elevati dovuto alla reazione specifica, cioè alla reale presenza di danno al DNA.

Per una corretta individuazione della quantità di spermatozoi danneggiati l'ultimo passaggio prevede di applicare gli stessi gate del campione di test anche al campione di controllo e valutare il grafico che mostra i valori di FITC dei due campioni contemporaneamente (vedi Figura 8).

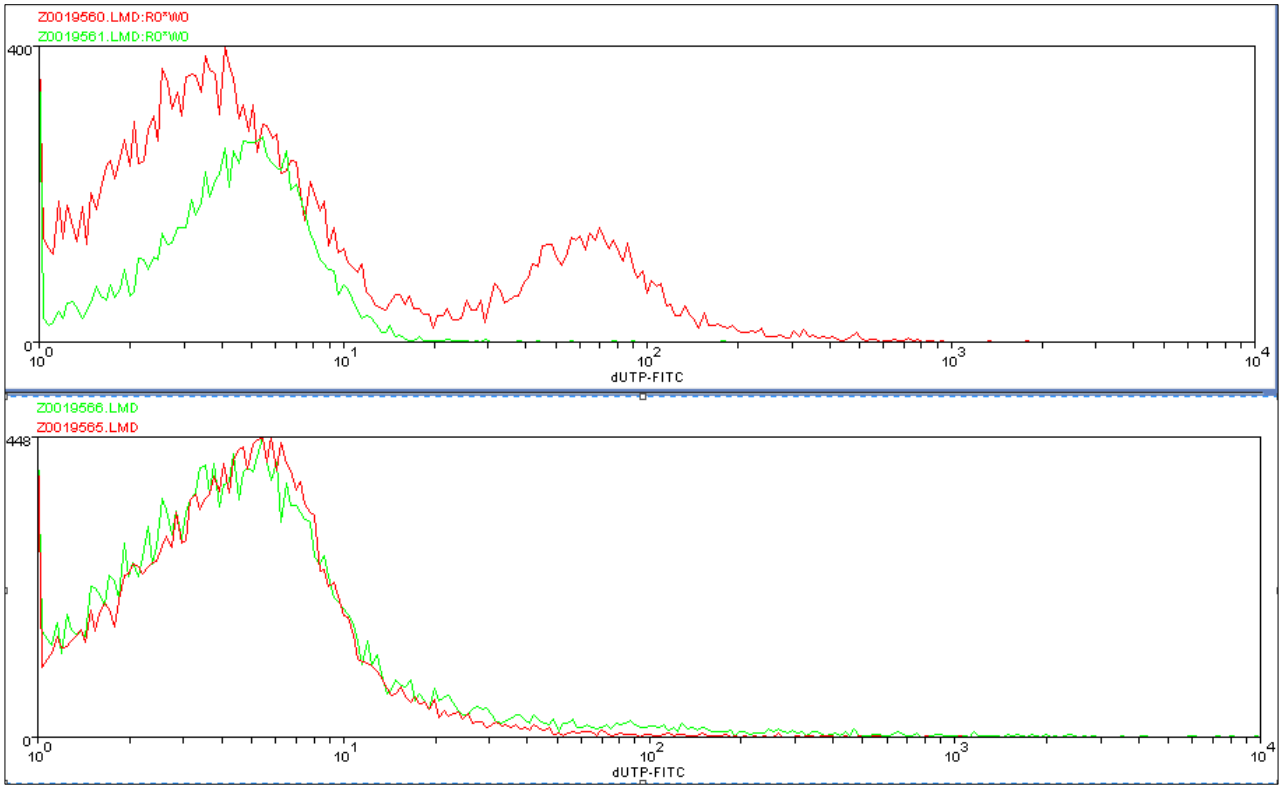


Figura 8: Istogrammi del danno al DNA del campione di controllo (colore rosso) e del campione di test (colore verde) relativi a due pazienti diversi.

Da un confronto di questo tipo ci si aspetta che la curva del campione di test copra quasi interamente la curva del campione di controllo perché nel test troveremo il colorante che si è legato in maniera aspecifica oltre al colorante rimasto legato dalla reazione specifica ai filamenti di DNA realmente danneggiati.

A questo punto bisognerebbe eseguire la sottrazione canale per canale del test dal campione. Ne risulterebbe l'istogramma di frequenze della reale quantità di danno al DNA degli spermatozoi.

Nella pratica il Prof. Bizzaro allinea i due grafici sul primo picco, esegue la sottrazione canale per canale e considera tutto ciò che rimane come danno al DNA. Altri ricercatori invece si limitano ad osservare che l'istogramma del danno al DNA è un istogramma bimodale nel quale la prima campana rappresenta sempre il controllo negativo (gli spermatozoi che si sono colorati in modo aspecifico) e la seconda campana rappresenta gli spermatozoi che davvero sono danneggiati. Da questa considerazione si deduce quanti spermatozoi sono danneggiati e quanto sono danneggiati valutando solo la seconda campana fino all'estremo destro dell'istogramma.

1.5.1 Il danno al DNA

Individuati gli spermatozoi l'esperto deve valutare il danno al DNA. Questa valutazione è importante perché determina il tipo di terapia da fare seguire al paziente. Ad esempio nel caso di un paziente con pochi danni al DNA è possibile iniziare subito una tecnica di fecondazione assistita come l'iniezione introcitoplasmatica, mentre per un paziente con molti danni al DNA è preferibile cercare di diminuire la quantità di danno, ad esempio con una dieta o uno stile di vita specifico, e solo successivamente tentare una tecnica di fecondazione assistita aumentando così le possibilità di buona riuscita.

La valutazione del danno al DNA è quindi molto importante e gli aspetti da prendere in considerazione sono la quantità di cellule danneggiate e l'entità del danno agli spermatozoi: un paziente con il 30% di spermatozoi poco danneggiati, cioè poco fluorescenti (Figura 9 – C2), ed un paziente con il 30% di spermatozoi danneggiati con un danno elevato, molto fluorescenti (Figura 9 – C1), pur avendo la stessa quantità di spermatozoi danneggiati dovranno essere individuati e trattati diversamente. Gli esperti hanno visto che non esiste una soglia biologica superata la quale un paziente sicuramente non è in grado di procreare e prima della quale sicuramente è fertile, però la quantità di danno e l'intensità dello stesso influiscono sulla fertilità.

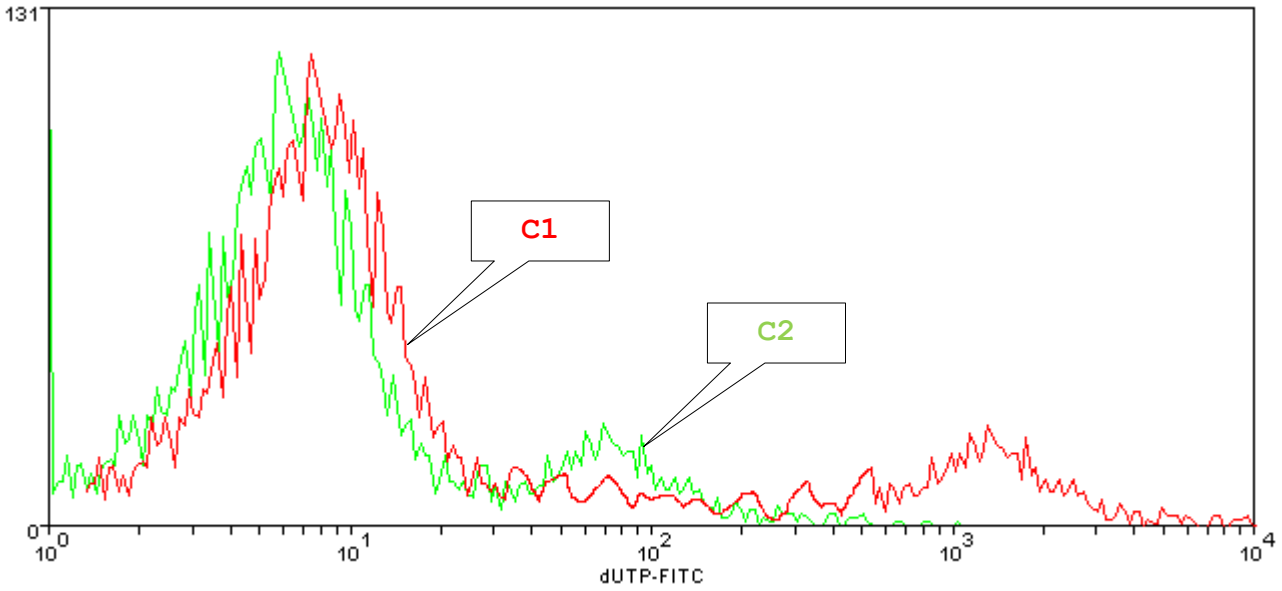


Figura 9: Danno al DNA degli spermatozoi in due diversi pazienti

Ad oggi le analisi vengono effettuate dall'esperto identificando il valore di questa soglia nel grafico del danno al DNA, sulla base della sua conoscenza ed esperienza e andando successivamente a calcolarne la percentuale di danneggiati rispetto al numero totale di spermatozoi.

Cap. 2 - I dati

Il citofluorimetro dopo l'elaborazione del campione salva le informazioni su un file del tipo "Z#######.LMD" in cui la sequenza di "#" rappresenta il numero progressivo del campione analizzato.
I dati risultanti dall'esperimento sono memorizzati nel file in formato binario secondo lo standard FCS.2.

2.1 Standard FCS2.0

Lo standard FCS2.0 *[5]* per i file di dati provenienti da citofluorimetri a flusso è basato sullo standard proposto da Murphy e Chused, FSC1.0 *[6]*.
Questo standard permette la descrizione dettagliata dell'esperimento includendo informazioni riguardanti lo strumento adottato, i parametri misurati, i valori dei parametri per ogni oggetto analizzato ed i risultati dell'analisi.

Il file può contenere uno o più analisi di campioni (data set). Ogni data set è suddiviso in almeno quattro parti obbligatorie: HEADER, TEXT, DATA e ANALYSIS. L'ordine di queste sezioni è arbitrario eccetto per la sezione HEADER che deve essere sempre la prima poiché contiene i puntatori ai byte di inizio e fine delle altre sezioni.
La sezione TEXT è composta da una serie di keyword e corrispettivi valori. Queste possono essere in numero e lunghezza qualsiasi (numero di byte). Le keyword che iniziano con il carattere '$' sono considerate standard FCS, a differenza delle keyword definite dall'utente le quali possono iniziare con un carattere qualsiasi.
Lo standard inoltre prevede un certo numero di keyword obbligatorie necessarie per poter correttamente interpretare il file.
La sezione DATA contiene i valori delle misurazioni effettuate dal citofluorimetro memorizzati come interi senza segno, floating point a singola precisione, floating point a doppia precisione oppure come valori ASCII.
La sezione ANALYSIS tipicamente memorizza informazioni supplementari ottenute elaborando i dati acquisiti.

Il file è strutturato come una sequenza di byte senza interruzioni, linee o formattazioni particolari. I dati nelle sezioni HEADER, TEXT e ANALYSIS sono scritti in formato ASCII.

2.2 Natura dei dati

Ogni campione di eiaculato di un paziente contiene mediamente dalle 20000 alle 40000 cellule di cui solitamente solo 5000 - 10000 sono spermatozoi.

Il campione nella pratica è composto da due file differenti:

- un file con il valore della keyword **$SRC** con un identificativo del paziente seguito dal segno "-" (es: $SRC!61g-!). Questo rappresenta il campione di controllo, cioè il campione in cui non è avvenuta la reazione che colora i danni al DNA (reazione aspecifica).
- un file, solitamente il successivo in ordine lessicografico, con la keyword **$SRC** valorizzata con lo stesso identificativo del paziente senza però essere seguito dal segno "-", oppure seguito dal segno "+" (es: $SRC!61g+!). Questo file rappresenta il risultato dell'analisi del campione in cui è avvenuta la reazione specifica per colorare il danno al DNA (campione di test).

2.2.1 Parametri

Indifferentemente dal fatto che si analizzi il campione di controllo o quello di test, il citofluorimetro misura 8 parametri per ogni cellula che lo attraversa:

- Parametro 1 = FL4 LOG → PI Log: misura logaritmica del valore di fluorescenza emessa dalla molecola di Propidio Ioduro
- Parametro 2 = FL1 LOG → dUTP-FITC: misura logaritmica del valore di fluorescenza emessa dalla molecola di Isotiocianato di Fluoresceina
- Parametro 3 = SS LOG → SS LOG: parametro fisico Side Scatter in misura logaritmica
- Parametro 4 = FS LOG → FS LOG: parametro fisico Forward Scatter in misura logaritmica
- Parametro 5 = FS → FS: parametro fisico Forward Scatter in misura lineare

- Parametro 6 = SS → SS: parametro fisico Side Scatter in misura lineare
- Parametro 7 = FL3 → PI LIN: misura lineare del valore di fluorescenza emessa dalla molecola di Propidio Ioduro
- Parametro 8 = AUX(FL1) → AUX(FL1): parametro ausiliario associato al parametro FL1 che rileva quindi la fluorescenza emessa dalla molecola di Isotiocianato di Fluoresceina con valori di gain e voltaggio differenti.

2.2.2 Scala

I valori dei parametri misurati dal citofluorimetro possono essere registrati sia in scala lineare che logaritmica. Solitamente la rappresentazione lineare si utilizza quando i valori hanno un range dinamico piuttosto ristretto (ad esempio quando si misura il contenuto di DNA di una popolazione cellulare), mentre si utilizza una rappresentazione logaritmica nei casi in cui il segnale da registrare presenta un range dinamico elevato come accade nella misura del danno al DNA *[2][3]*.

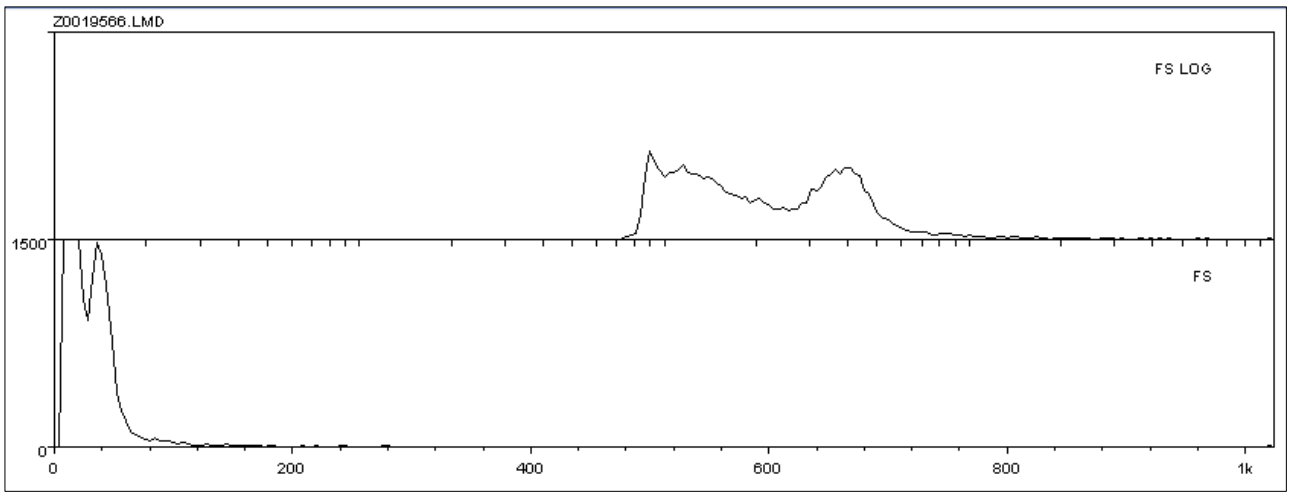


Figura 10: Visualizzazione della *feature* FS registrata in scala logaritmica e lineare

I valori di alcune *feature* sono memorizzati sia in forma lineare che logaritmica per poter scegliere successivamente la rappresentazione che permette di eseguire una analisi più efficace.

2.2.3 Valori ammessi

I data set non presentano valori mancanti o fuori scala. Tutte le *feature* misurate hanno valori numerici interi compresi nel range [0..1023]. Questi valori non hanno una unità di misura costante. Questo significa che se si stanno osservando dei virus ed un oggetto del campione misura SS=600, l'unità di misura da adottare saranno i nanometri, mentre se si sta analizzando un eiaculato lo stesso valore di SS=600 corrisponde all'unità di misura dei micron. La stessa considerazione è necessaria per la misurazione delle sostanze fluorescenti, quindi, prima di eseguire l'analisi dei campioni, è necessario tarare lo strumento sulla base degli oggetti che si vorranno osservare, misurando delle sfere di dimensione nota e con una quantità di molecole fluorescenti nota. Dopo la taratura sarà possibile risalire alle unità di misura appropriate per ogni parametro.

Il citofluorimetro viene impostato[11] per non memorizzare informazioni su particelle più piccole di una certa dimensione perché gli esperti sanno a priori che si tratterebbe di particelle o pezzi di cellule che non influenzerebbero l'analisi degli spermatozoi.
Questo è possibile verificarlo in Figura 11 visualizzando il dot plot dei valori di FS Log e SS Log e notando che non sono presenti punti al di sotto di un determinato valore di FS Log.

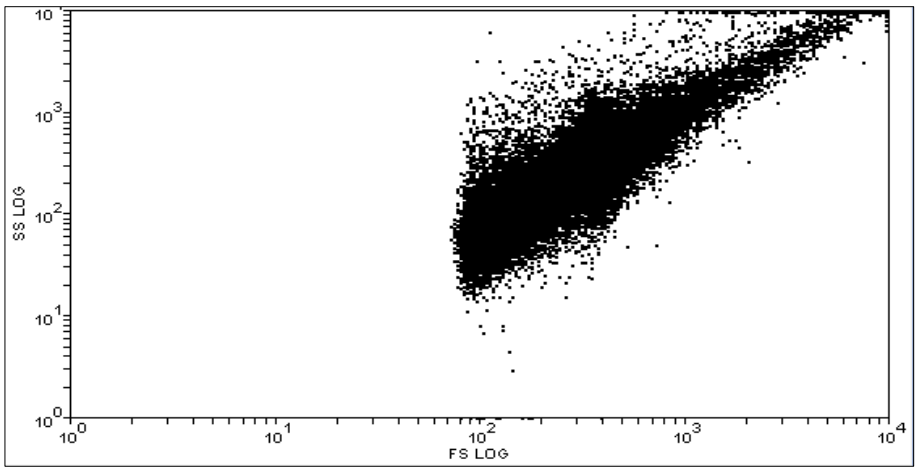


Figura 11: Valori minimi registrati per il parametro FSLog (asse delle ascisse)

[11] Per tarare il valore di FS sotto il quale il citofluorimetro non deve registrare i risultati si esegue la seguente procedura:

- Si diluisce l'eiaculato con un liquido opportunamente filtrato (acqua di trasporto) per poter analizzare il campione al citofluorimetro.
- Si analizza al citofluorimetro solo il liquido di trasporto e si annota il valore maggiore di FS Log che viene rilevato.
- Il valore di FS Log precedentemente trovato viene utilizzato come soglia per poter discriminare le cellule dell'eiaculato da quelle del liquido di trasporto che non devono essere analizzate.

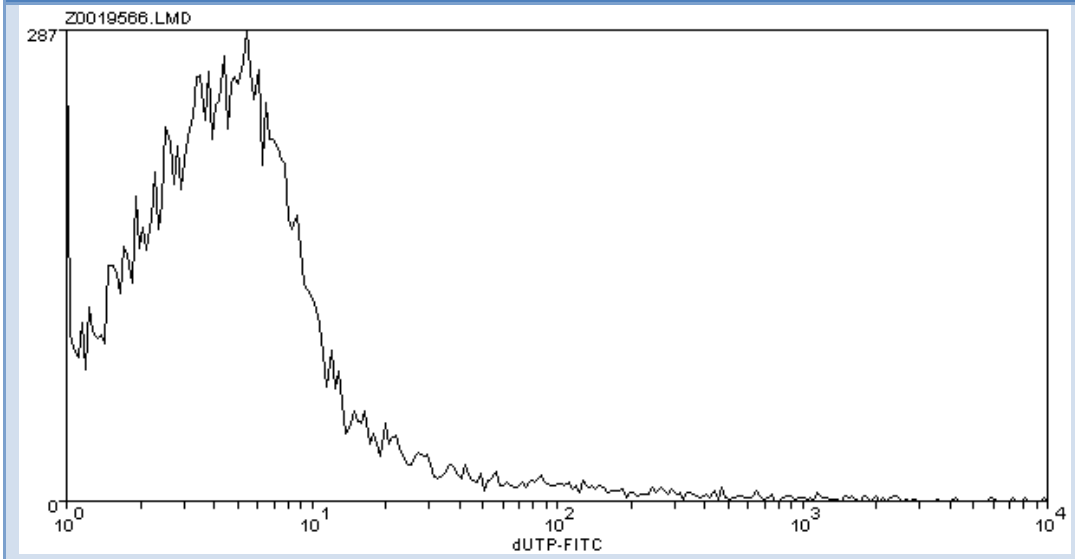
2.2.4 Data set

I data set a disposizione sono stati ottenuti con diverse impostazioni del citofluorimetro per quanto riguarda i valori di gain e voltaggio dei parametri misurati. Inoltre, come spiegato precedentemente, un campione di un paziente è suddiviso tra test e controllo. Non potendo garantire in assoluto che l'attivazione della reazione chimica per marcare il danno al DNA non influenzi la misurazione degli altri parametri, bisogna concludere che data set di controllo e di test, anche se appartenenti allo stesso campione, non sono omogenei.
Nella scelta del data set da utilizzare per testare gli algoritmi di Data Mining si è preferito orientarsi verso i campioni in cui è avvenuta la reazione specifica. Questa scelta è stata motivata dal fatto che il campione di test è quello in cui l'esperto del dominio vuole individuare gli spermatozoi e successivamente valutarne il danno.
Per quanto riguarda invece la scelta del voltaggio e gain, non ci si è orientati verso dei valori predeterminati, ma si sono utilizzati quelli corrispondenti ai data set più numerosi.

Nel seguito sono riportati i parametri statistici[12] più importanti del data set utilizzato per testare gli algoritmi di Data Mining.
Il data set fa riferimento al campione 68i+ Z0019566 in cui sono presenti le misurazioni di 21106 cellule. L'etichettatura manuale da parte dell'esperto (gate su FSLog-SSLog e successivo gate su PI) ha evidenziato la presenza di 5891 spermatozoi (etichetta SPERMATOZOO) e 15215 cellule di altro genere (etichetta UNKNOWN).

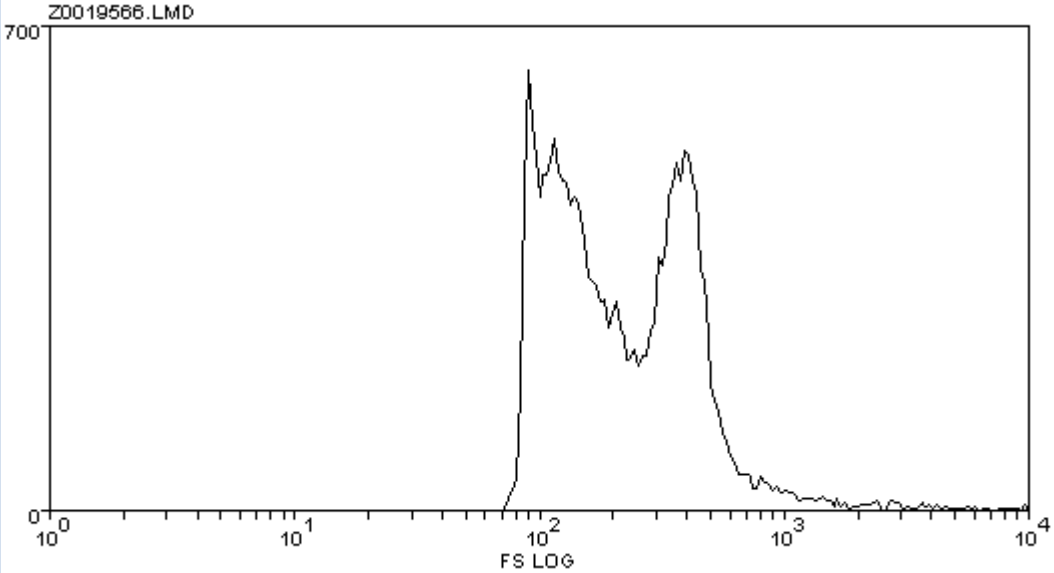
[12] Le immagini sono state prodotte con il software Weasel; i dati statistici sono stati calcolati dal software di data mining Weka.

dUTP-FITC



Valori mancanti	0 (0%)
Valori distinti	716
Valori unici	136 (1%)
Minimo	0
Massimo	1023
Valor medio	122,528
Deviazione standard	130,077

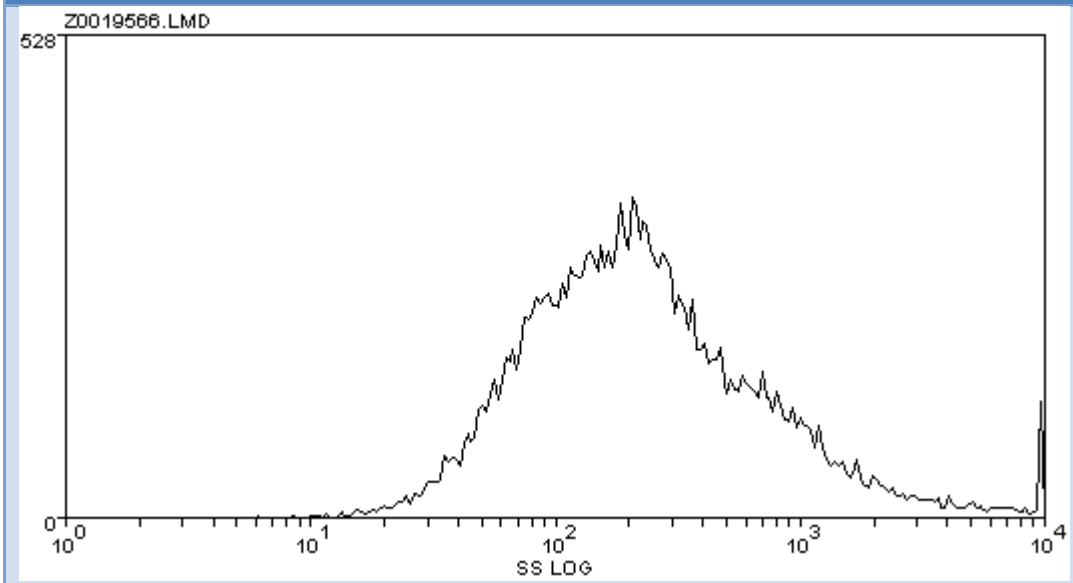
FSLog



Valori mancanti	0 (0%)

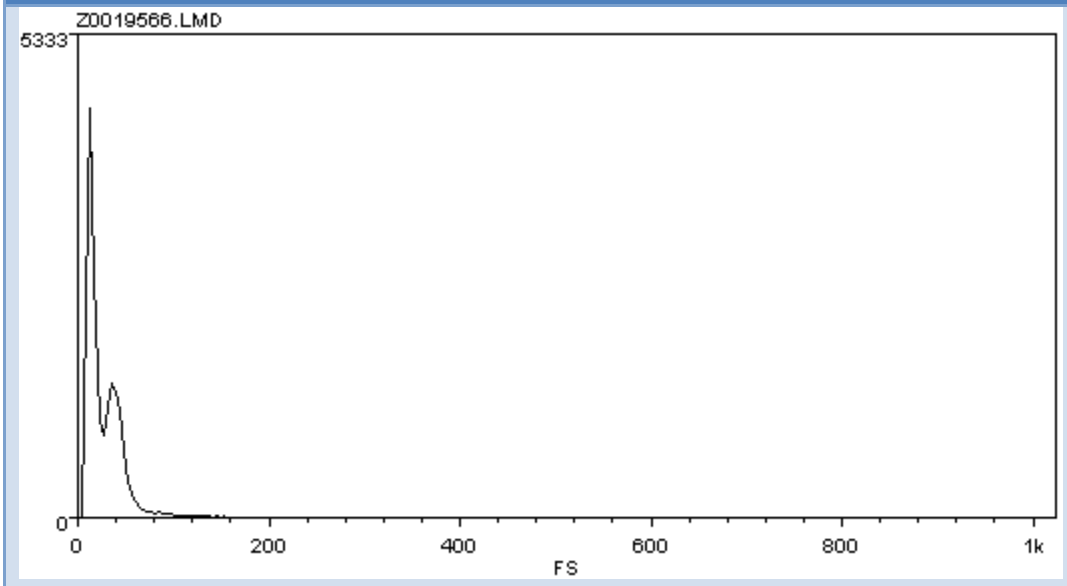
Valori distinti	482
Valori unici	62 (0%)
Minimo	472
Massimo	1023
Valor medio	599,239
Deviazione standard	78,657

SSLog



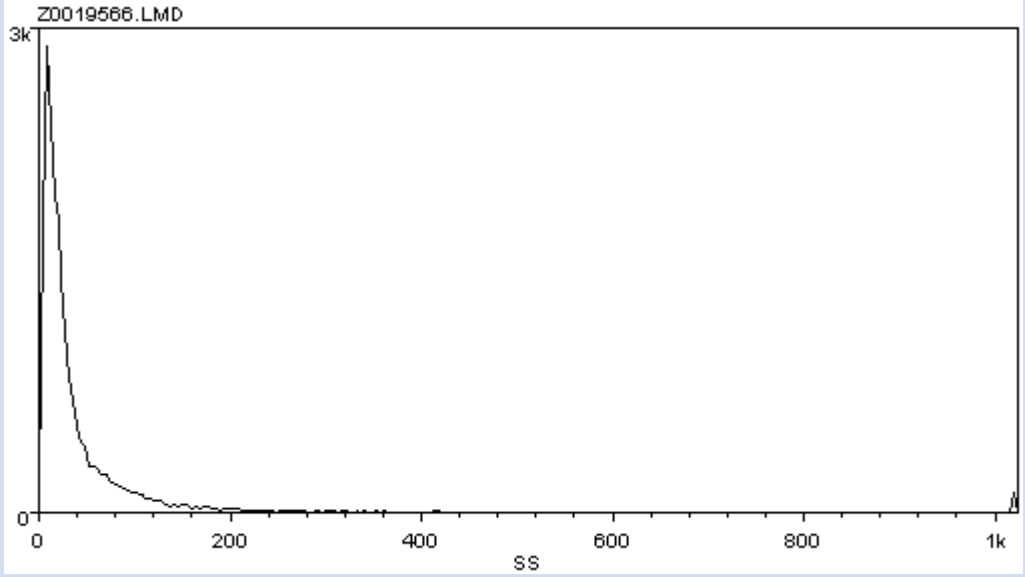
Valori mancanti	0 (0%)
Valori distinti	739
Valori unici	54 (0%)
Minimo	0
Massimo	1023
Valor medio	603,817
Deviazione standard	124,029

FS



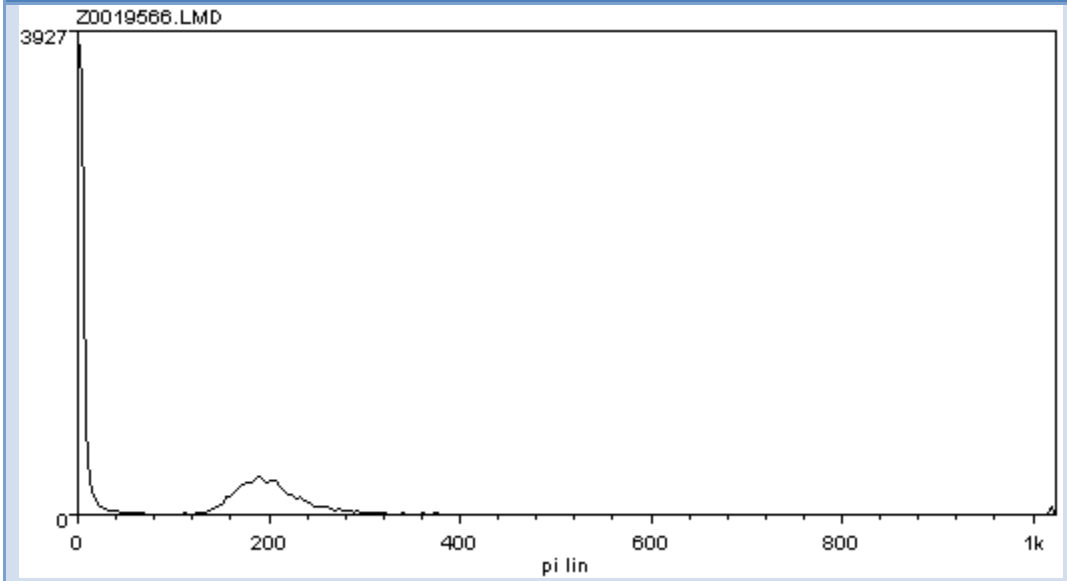
Valori mancanti	0 (0%)
Valori distinti	357
Valori unici	150 (1%)
Minimo	7
Massimo	1023
Valor medio	31,119
Deviazione standard	51,607

SS

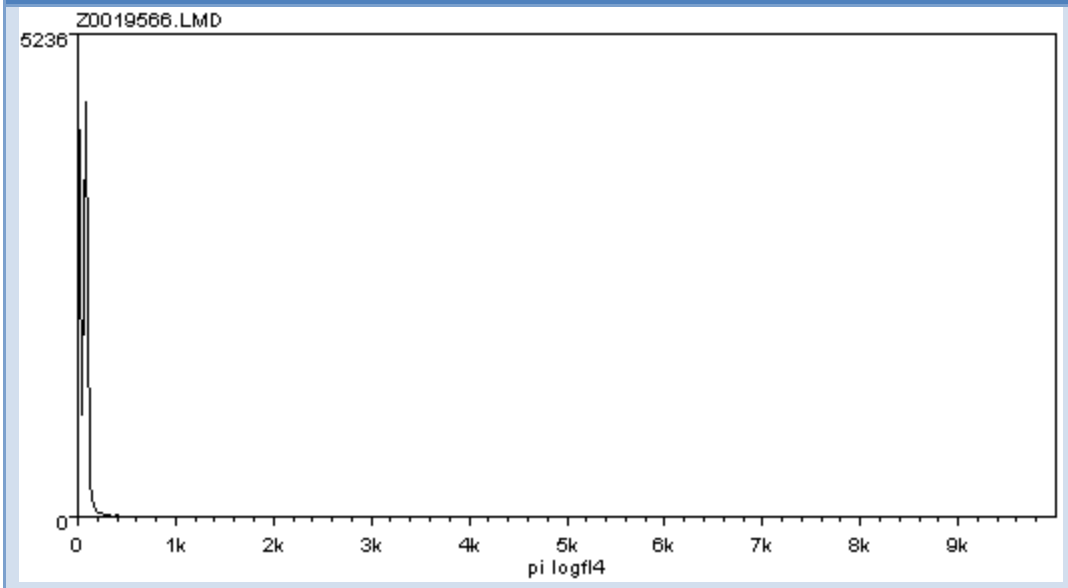


Valori mancanti	0 (0%)
Valori distinti	583
Valori unici	205 (1%)
Minimo	0
Massimo	1023
Valor medio	50,84
Deviazione standard	112,119

PI



Valori mancanti	0 (0%)
Valori distinti	545
Valori unici	176 (1%)
Minimo	0
Massimo	1023
Valor medio	68,69
Deviazione standard	121,652

PILog	
	
Valori mancanti	0 (0%)
Valori distinti	668
Valori unici	108 (1%)
Minimo	0
Massimo	932
Valor medio	157,376
Deviazione standard	216,572

2.3 Estrazione dei dati

Per poter individuare correttamente gli spermatozoi all'interno del campione è stato necessario sviluppare un insieme di software che fossero in grado di interpretare il file in output dal citofluorimetro per estrarre i parametri di ogni singola cellula e successivamente analizzarli con tecniche di Data Mining.

Il file prodotto dal citofluorimetro è interpretato da un convertitore che ne esegue il parsing identificando le cellule del campione con i relativi parametri misurati e li memorizza nel formato ARFF per il successivo utilizzo con il software Weka.

I file ARFF (Attribute-Relation File Format) sono file di testo ASCII che descrivono una lista di istanze le quali condividono una serie di attributi.

Un file ARFF è composto da due diverse sezione: HEADER e DATA.
Nella sezione HEADER è presente il nome della relazione ed una lista ordinata di attributi con il relativo tipo (es: numerico, nominale, data).
Nela sezione DATA ogni riga rappresenta un'istanza, ed ogni colonna corrisponde agli attributi secondo l'ordinamento definito nella sezione HEADER.

Nel contesto di dati citofluorimetri le istanze rappresentano le cellule del campione mentre gli attributi rappresentano le *feature* delle cellule. L'incrocio tra una istanza e un attributo rappresenta il valore della *feature* misurata dal citofluorimetro relativamente ad una cellula del campione.

2.4 Analisi dei dati

Una volta estratti i dati e collezionati in un formato idoneo all'analisi, il lavoro si è concentrato sull'individuazione di un metodo operativo che riuscisse ad isolare nel campione le sole cellule di interesse: gli spermatozoi.
A questo scopo si sono utilizzati algoritmi appartenenti alle due grandi famiglie di tecniche di Machine Learning: tecniche supervisionate e tecniche non supervisionate.
Le tecniche non supervisionate prendono in considerazione tutti gli attributi portando alla luce pattern e strutture caratteristiche del data set non avendo a disposizione una classificazione delle istanze. Le tecniche supervisionate, come la classificazione, al contrario si basano sulla presenza di un attributo, la classe, che etichetta le diverse istanze del data set. Un algoritmo supervisionato, basandosi sugli esempi a disposizione, cerca di individuare le relazioni che legano il valore dell'etichetta della istanza con il valore degli altri attributi (attributi predittori). In pratica la classificazione cerca di stimare o predire il valore di un attributo target.

Una volta applicate le tecniche di Data Mining è importante riuscire ad individuare quale è stata la migliore e come i suoi parametri caratteristici hanno influenzato sul risultato. Sia per le tecniche non supervisionate che per quelle supervisionate si è ricorso a metodi di valutazione esterni *[9]*. Alcuni data set a disposizione sono infatti corredati da un ulteriore attributo, la classe, che etichetta le singole istanze come spermatozoi o non-spermatozoi. La

valutazione del risultato è avvenuta confrontando il grado di corrispondenza[13] tra l'etichetta del cluster a cui è stata assegnata l'istanza e l'etichetta di classe attribuita dall'esperto del dominio. Per fare questo si sono valutate alcune misure, mutuate dalla classificazione:

- Entropia: misura del disordine all'interno del cluster. Per ogni cluster *j* viene calcolata la distribuzione di classe p_{ij}, viene cioè calcolata la probabilità che un oggetto del cluster *i* appartenga alla classe *j* come $p_{ij} = m_{ij}/m_i$ dove m_i è il numero di oggetti nel cluster i e m_{ij} è il numero di oggetti di classe j nel cluster i. L'entropia del cluster *i* è calcolata usando la formula $e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij}$ dove L è il numero di classi. L'entropia totale del clustering è calcolata come la somma delle entropie di ogni cluster pesata con la dimensione del cluster, cioè, $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$ dove K è il numero di cluster individuati e *m* è il numero totale di punti nel data set.
- Purezza: indica la massima probabilità di classe all'interno di un cluster. La purezza del cluster i è $p_i = \max_j p_{ij}$, mentre la purezza dell'intero clustering è $purezza = \sum_{i=1}^{K} \frac{m_i}{m} p_i$.
- Precisione: indica la probabilità che un oggetto in un cluster appartenga ad una certa classe. La precisione del cluster *i* rispetto alla classe *j* è $precisione(i,j) = p_{ij}$.
- Recall: misura la frazione di oggetti di una classe contenuta in un cluster. Il valore di recall del cluster *i* rispetto alla classe *j* è $recall(i,j) = m_{ij}/m_j$ dove m_j è il numero di oggetti di classe *j*.
- F-measure: combinazione di precisione e recall che misura quanto il cluster contiene tutti e nient'altro che oggetti di una classe. F-measure di un cluster *i* rispetto alla classe *j* è $F(i,j) = \big(2 * precision(i,j) * recall(i,j)\big)/(precision(i,j) * recall(i,j))$.

In aggiunta a queste misure per valutare le tecniche di classificazione sono state prese in esame anche:

- Confusion Matrix che mette in evidenza il risultato della classificazione rispetto alla classe reale delle istanze. Dalla confusion matrix si sono calcolate le metriche:

[13] Nelle tabelle riassuntive degli esperimenti per ogni cluster individuato dall'algoritmo sarà indicato il numero di cellule etichettate dall'esperto del dominio come spermatozoi, "S", e le cellule etichettate come non-spermatozoi (unknown), "U".

- Accuratezza = (TP + TN) / (TP + TN + FP + FN)
- Sensitività = TP / (TP / FN)
- Specificità = TN / (TN + FP)
- Precisione = TP / (TP + FP)
- TP rate = TP / (TP + FN) *100
- FP rate = FP / (FP + TN) * 100

Dove TP sono gli spermatozoi classificati come tali, TN sono non-spermatozoi classificati come tali, FP sono non-spermatozoi classificati come spermatozoi, FN sonno spermatozoi classificato come non-spermatozoi.

- K-statistic che misura il miglioramento del modello di classificazione rispetto al classificatore guidato dalle probabilità stimate.

2.5 *Feature* di maggiore interesse

Come spiegato nel paragrafo 2.2.2 le *feature* sono misurate dal citofluorimetro in doppia scala: lineare e logaritmica. Il primo approccio ai dati ha avuto come obiettivo l'individuazione della rappresentazione migliore tra le due. Questo è stato svolto con un algoritmo comunemente utilizzato durante la fase esplorativa di un data set: K-Means[14].

Sono state testate una ad una le singole *feature* e si sono confrontati i risultati, in termini di individuazione di spermatozoi, tra quelle che misurano la stessa proprietà fisica della cellula.

Individuate le rappresentazioni migliori si è proseguito il lavoro applicando gli algoritmi di data mining alle *feature* congiuntamente.

[14] Per una spiegazione più dettagliata consultare il paragrafo 3.1

2.5.1 *Feature* FS – FSLog

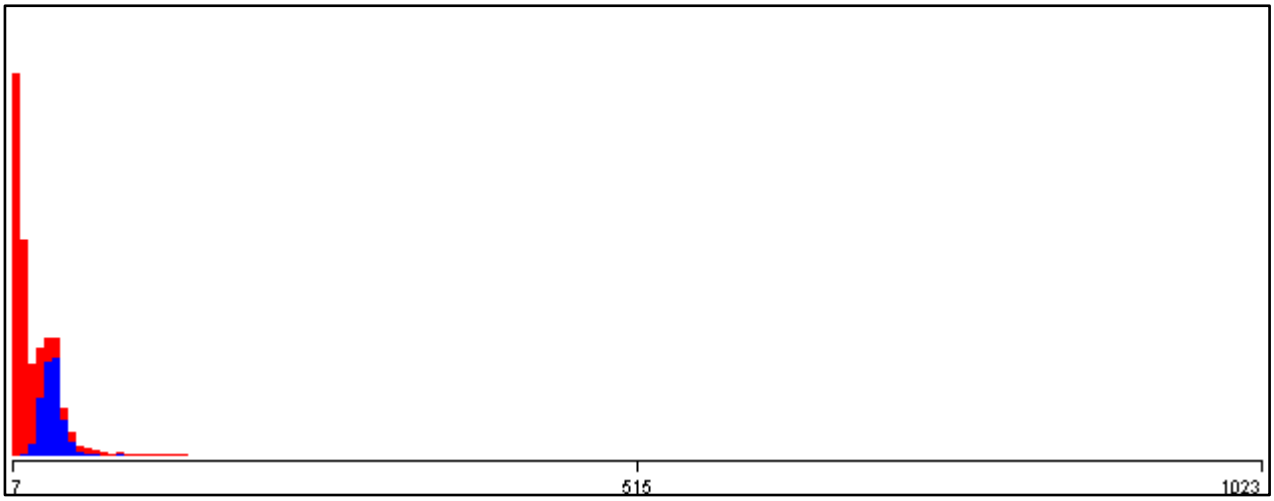


Figura 12: Istogramma di frequenze della *feature* FS. In blu gli spermatozoi, in rosso le cellule unknown

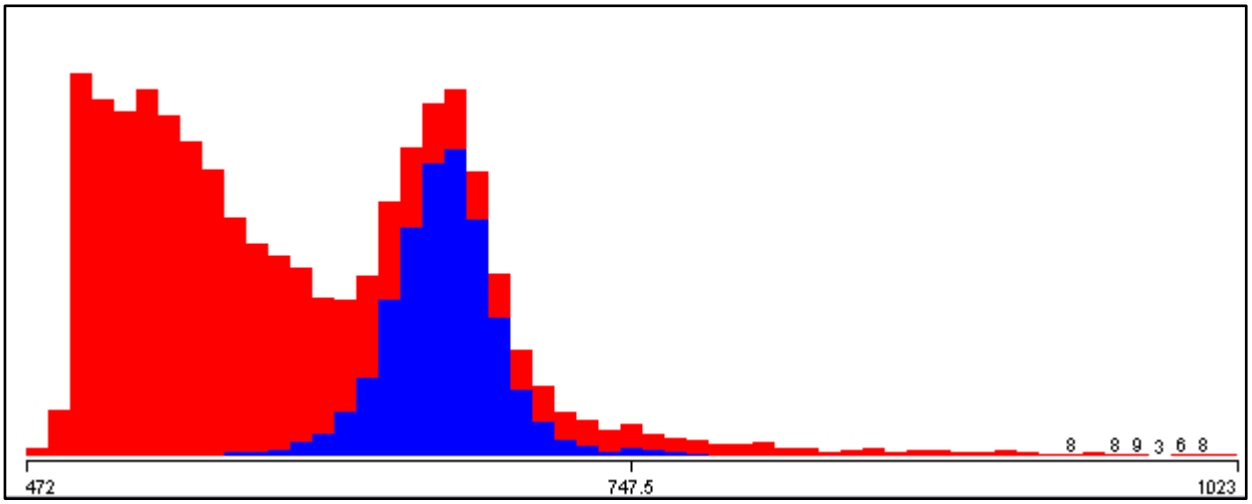


Figura 13: Istogramma di frequenze della *feature* FSLog. In blu gli spermatozoi, in rosso le cellule unknown

2.5.1.1 Funzione di distanza

Funzione di distanza		Entropia Cluster0	Entropia Cluster1	Purezza Cluster0		Purezza Cluster1		Recall Cluster0		Recall Cluster1	
				S	U	S	U	S	U	S	U
euclidean	FSLog	0,962	0,084	0,614			0,989	0,978	0,238	0,021	0,761
	FS	0	0,857		1		0,718	0	0,011	1	0,988
manhattan	FSLog	0,978	0,060	0,586			0,992	0,986	0,269	0,013	0,730
	FS	0,928	0,181	0,655			0,972	0,941	0,245	0,037	0,883

Considerazioni per FSLog: utilizzare come **distanceFunction** la distanza di Manhattan in alternativa alla distanza Euclidea non migliora il risultato del clustering: il Cluster1 continua ad essere ben rappresentato dagli unknown mentre il Cluster0 seppure contenga la maggior parte degli spermatozoi è caratterizzato da una alta entropia per cui presenta una mescolanza di spermatozoi e cellule unknown.

Considerazione per FS: utilizzare come **distanceFunction** la distanza di Manhattan rispetto alla distanza Euclidea migliora il risultato del clustering: ora è possibile individuare due cluster distinti di cui il Cluster1 è composto principalmente da cellule unknown e caratterizzato da una elevata purezza, ed il Cluster0 con un valore elevato di recall per quanto riguarda gli spermatozoi, però caratterizzato da una elevata entropia.

2.5.1.2 Numero di cluster

# di cluster		Entropia Clustering	Purezza Clustering
2	FSLog	0,475001	0,822468
	FS	0,85033	0,720885
3	FSLog	0,461767	0,830712
	FS	0,576723	0,842841
4	FSLog	0,423012	0,875438
	FS	0,496786	0,861556
5	FSLog	0,432091	0,850185
	FS	0,440543	0,874159
9	FSLog	0,386992	0,884772
	FS	0,407075	0,87757

Aumentare il valore di **numClusters** ha come effetto generale la diminuzione dell'entropia e l'aumento della purezza dello schema di clustering. Nel complesso però i risultati migliori si sono ottenuti con la *feature* FSLog. Osservano gli istogrammi di frequenza delle due

feature è possibile osservare come per FS, rispetto a FSLog, i dati siano molto compressi attorno al loro valor medio; probabilmente questa caratteristica è stata determinante nell'ottenere risultati peggiori rispetto alla versione logaritmica.

2.5.2 *Feature* SS – SSLog

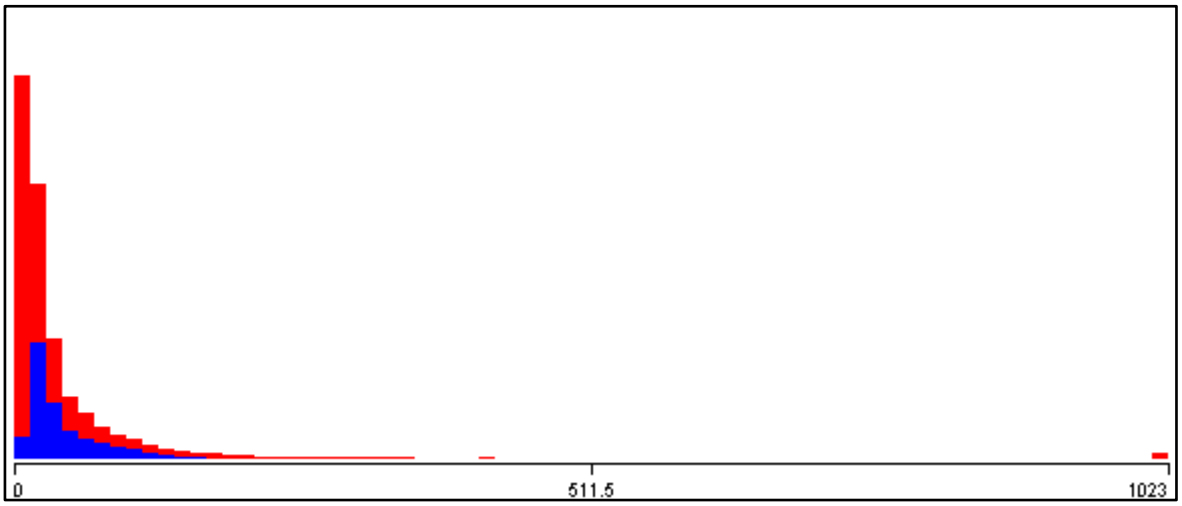


Figura 14: Istogramma di frequenze della *feature* SS. In blu gli spermatozoi, in rosso le cellule unknown

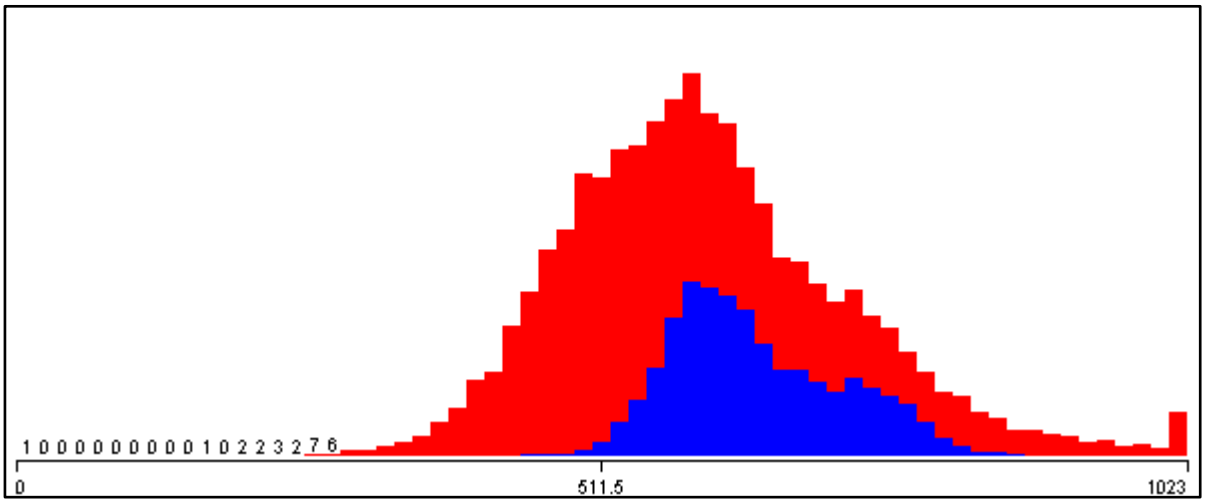


Figura 15: Istogramma di frequenze della *feature* SSLog. In blu gli spermatozoi, in rosso le cellule unknown

2.5.2.1 Funzione di distanza

Funzione di distanza		Entropia Cluster0	Entropia Cluster1	Purezza Cluster0		Purezza Cluster1		Recall Cluster0		Recall Cluster1	
				S	U	S	U	S	U	S	U
euclidean	SSLog	0,964	0,758		0,610		0,780	0,492	0,289	0,507	0,701
	SS	0	0,861		1		0,715	0	0,02	1	0,972
manhattan	SSLog	0,974	0,686		0,593		0,817	0,626	0,353	0,373	0,646
	SS	0,915	0,838		0,669		0,732	0,216	0,170	0,783	0,829

Considerazione per SSLog: utilizzare come **distanceFunction** la distanza di Manhattan in alternativa alla distanza Euclidea non migliora sensibilmente il risultato del clustering che continua ad essere caratterizzato da una entropia globale elevata e recall degli spermatozoi nei singoli cluster basso.

Considerazioni per SS: utilizzare come **distanceFunction** la distanza di Manhattan rispetto alla distanza Euclidea non migliora il risultato del clustering: ora è possibile individuare due cluster distinti ma entrambi presentano una alta entropia ed un basso valore di purezza.

2.5.2.2 Numero di cluster

# di cluster		Entropia Clustering	Purezza Clustering
2	SSLog	0,831097	0,720885
	SS	0,844917	0,720885
3	SSLog	0,72947	0,720885
	SS	0,842208	0,720885
4	SSLog	0,712398	0,720885
	SS	0,836852	0,720885
5	SSLog	0,703942	0,720885
	SS	0,824289	0,720885
9	SSLog	0,685731	0,720885
	SS	0,740048	0,720885

Aumentare il valore di **numClusters** ha come effetto generale la diminuzione dell'entropia dello schema di clustering, mentre la purezza rimane stabile. Nel complesso però i risultati migliori si sono ottenuti con la *feature* SSLog. Osservano gli istogrammi di frequenza delle due *feature* è possibile osservare come per SS, rispetto a SSLog, i dati siano molto compressi attorno al loro valor medio; probabilmente questa caratteristica è stata determinante nell'ottenere risultati peggiori rispetto alla versione logaritmica.

2.5.3 *Feature* PILog -PI

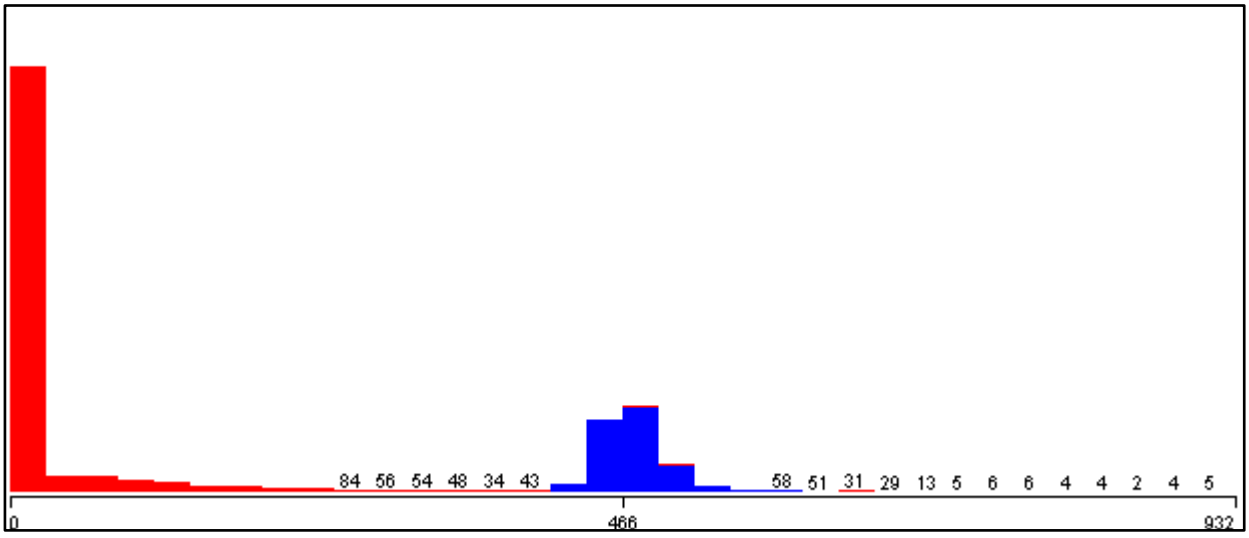


Figura 16: Istogramma di frequenze della *feature* PILog. In blu gli spermatozoi, in rosso le cellule unknown

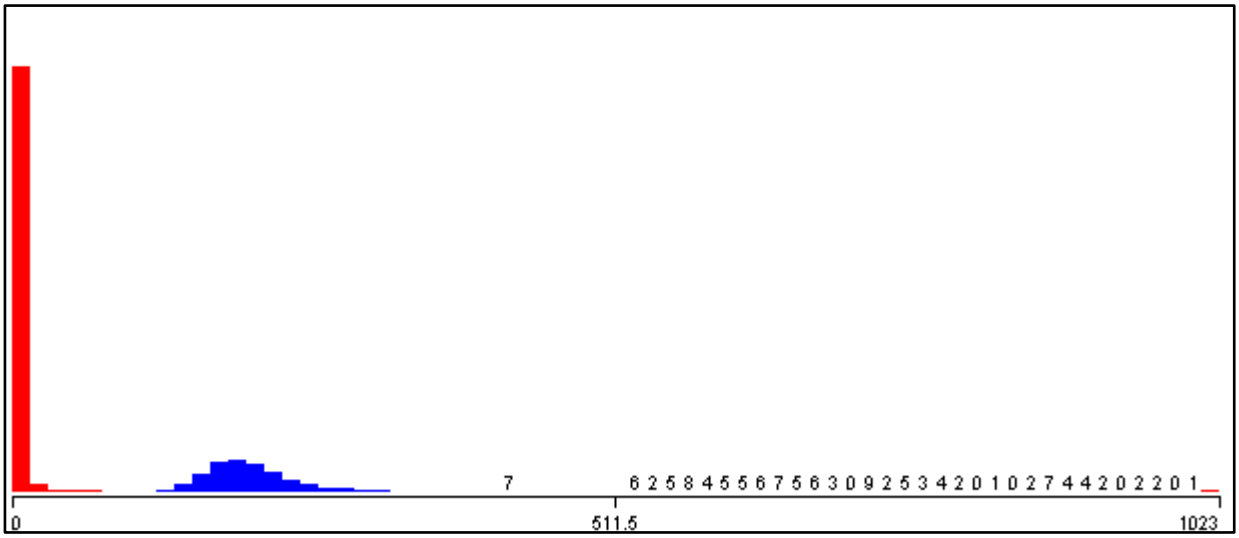


Figura 17: Istogramma di frequenze della *feature* PI. In blu gli spermatozoi, in rosso le cellule unknown

2.5.3.1 Funzione di distanza

Funzione di distanza		Entropia Cluster0	Entropia Cluster1	Purezza Cluster0		Purezza Cluster1		Recall Cluster0		Recall Cluster1	
				S	U	S	U	S	U	S	U
euclidean	PI	0,279	0	0,951			1	1	0,019	0	0,980
	PILog	0,463	0	0,901			1	1	0,042	0	0,957
manhattan	PI	0,293	0	0,948			1	1	0,021	0	0,978
	PILog	0,477	0	0,897			1	1	0,044	0	1

Considerazione per PI e PILog: utilizzare come **distanceFunction** la distanza di Manhattan o la distanza Euclidea non cambia sostanzialmente il risultato del clustering: entrambi i cluster continuano ad essere caratterizzati da bassa entropia e alta purezza.

2.5.3.2 Numero di cluster

# di cluster		Entropia Clustering	Purezza Clustering
2	PI	0,081869	0,985833
	PILog	0,143623	0,969535
3	PI	0,064025	0,988534
	PILog	0,132469	0,968871
4	PI	0,060553	0,98896
	PILog	0,094629	0,979011
5	PI	0,064155	0,988297
	PILog	0,098757	0,980906
9	PI	0,055758	0,989624
	PILog	0,07227	0,983891

Aumentando il valore di **numClusters** la purezza dello schema di clustering migliora e diminuisce l'entropia. Sia PI che PILog contribuiscono efficacemente ad individuare gli spermatozoi, PI però garantisce una diminuzione dell'entropia maggiore.

Cap. 3 - Clustering

La tecnica non supervisionata di data mining più comune è il clustering *[7][8]*. Il clustering concerne il raggruppamento di oggetti, le istanze del data set, in classi che hanno caratteristiche in comune. Ogni raggruppamento, chiamato cluster, contiene tutti gli oggetti che sono simili tra solo e differenti dagli oggetti appartenenti ad altri cluster. Il clustering infatti suddivide gli oggetti del data set in gruppi il più possibile omogenei cercando di massimizzare la similarità intra-gruppo e minimizzare la similarità inter-gruppo senza avere come obiettivo la previsione o la stima di un certo attributo.

Tra le diverse tecniche di clustering *[9]* sono stati presi in esame algoritmi basati su concetti molto diversi l'uno dall'altro per cercare di individuare la tecnica che conferisce i migliori risultati nell'individuazione degli spermatozoi. A tal proposito sono stati utilizzati metodi partitivi, basati sulla densità e su modelli parametrici.

Ogni algoritmo di clustering è stato provato più volte sul data set variando in ogni esperimento i parametri caratteristici dello specifico algoritmo al fine di determinare il migliore settaggio.

Gli attributi presi in esame sono stati:

- FSLog
- SSLog
- FS
- SS
- PI
- PIL

Non è invece stata trattata la *feature* AUX in quanto correlata ad una *feature*, variabile da esperimento ad esperimento, tra quelle prese in esame. Per quanto riguarda la *feature* FITC, che indica il danno al DNA, anche questa non è stata presa in esame in tale fase del progetto in quanto bisognerebbe riuscire a garantire che spermatozoi danneggiati non venissero assimilati a cellule non-spermatozoi. È impossibile però garantire questo fatto senza sapere qual è la "normale" quantità di DNA danneggiato in quel tipo di cellula, e questo è possibile

saperlo solo se si conosce il tipo di cellula in esame, che è proprio l'output che vogliamo dal clustering.

I test sono stati effettuati sul campione 68i+ Z0019566 etichettato manualmente dall'esperto (gate su FSLog/SSLog e successivamente gate su PI).

3.1 K-Means

K-Means è una tecnica di clustering partitiva[15] basata su prototipi[16], la quale cerca di individuare un numero di cluster K rappresentati dai loro centroidi.
L'algoritmo procede secondo questi passi:

- Fase 1: si imposta il numero K di cluster con cui si vuole sia suddiviso il data set.
- Fase 2: in modo random vengono scelte K istanze del data set che saranno i centri iniziali dei cluster.
- Fase 3: ogni istanza individua il centro del cluster più vicino secondo una metrica predefinita. In questo modo ogni centro diventa il possessore di un certo numero di istanze partizionando l'insieme in K cluster diversi.
- Fase 4: ogni cluster individua il suo centroide, il suo baricentro, che diventa il nuovo centro del cluster.
- Fase 5: si ripetono le fasi da 3 a 5 fino a quando i centroidi non mi modificano più (convergenza) oppure si raggiunge il numero massimo di iterazioni previste.

L'algoritmo K-Means è stato applicato al data set variando il numero di cluster **numClusters,** il valore del **seed** iniziale, il numero massimo di iterazioni **maxIterations** e la funzione di distanza **distanceFunction**. Dovendo cercare uno schema di clustering che distingua gli spermatozoi dalle cellule di non-spermatozoi, il minor numero di cluster cercati è pari a 2. A posteriori si è visto che il numero massimo d'iterazioni per ogni esperimento non è mai arrivato al valore di **maxIterations** settato, quindi non sono stati riportati i risultati per diversi valori del parametro **maxIterations**. Per quanto riguarda il

[15] Le tecniche partitive dividono il data set in cluster non sovrapposti, cioè ogni istanza è assegnata ad un unico cluster.
[16] Nel clustering basato su prototipi un cluster è definito come un insieme di oggetti simili, secondo una qualche misura, ad un rappresentante che viene chiamato centroide o medoide nel caso di attributi categorici.

seed iniziale non sono state riportate le prove effettuate poiché variandone il valore lo schema di clustering risultante non ha mai presentato significative differenze.

3.1.1 *Feature* FSLog - SSLog - PI

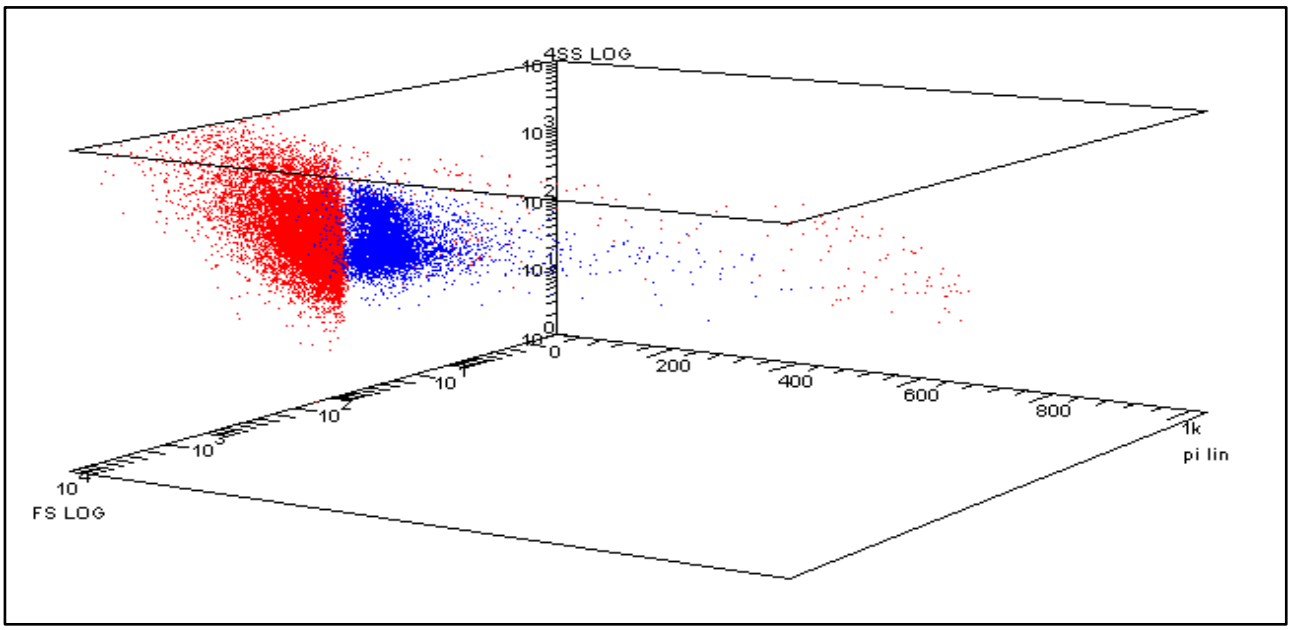


Figura 18: 3-D dot plot FSLog-SSLog-PI delle cellule del campione. In blu gli spermatozoi, in rosso le cellule unknown

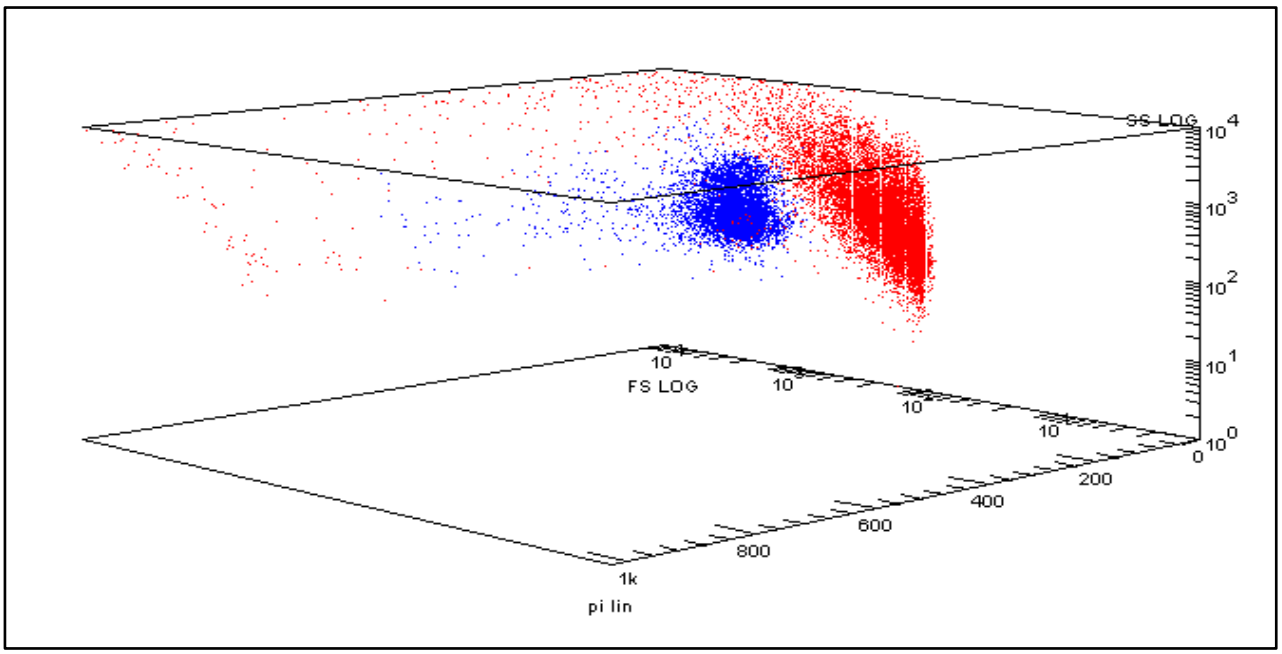


Figura 19: 3-D dot plot FSLog-SSLog-PI delle cellule del campione da un'altra angolazione

3.1.1.1 Funzione di distanza

Funzione di distanza	Entropia Cluster0	Entropia Cluster1	Purezza Cluster0		Purezza Cluster1		Recall Cluster0		Recall Cluster1	
			S	U	S	U	S	U	S	U
euclidean	0,916	0,018	0,668	0,332	0,002	0,998	0,996	0,191	0,003	0,808
manhattan	0,887	0,011	0,695	0,305	0,123	0,877	0,997	0,169	0,002	0,830

Utilizzare come **distanceFunction** la distanza di Manhattan o la distanza Euclidea non cambia sostanzialmente il risultato del clustering: l'entropia del cluster degli spermatozoi diminuisce leggermente ma continua ad avere un valore troppo elevato.

3.1.1.2 Numero di cluster

# cluster	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6	C7	C8
2	0,391571	0,861272	S	5870	21							
			U	2907	12308							
3	0,096479	0,985075	S	5804	7	80						
			U	228	11533	3454						
4	0,050567	0,992561	S	5832	1	56	2					
			U	98	5016	882	9219					
5	0,051458	0,99114	S	64	5772	53	0	2				
			U	164	68	1448	8146	5389				
6	0,049493	0,99204	S	2139	3699	44	0	0	9			
			U	27	88	159	8126	5377	1438			
7	0,034086	0,994693	S	2144	3692	40	0	1	1	13		
			U	14	43	145	6190	5625	2526	672		
8	0,027181	0,995641	S	2128	3722	39	0	1	11	0	0	
			U	9	32	136	6454	3691	499	1949	2445	
9	0,024573	0,996494	S	185	3679	4	0	1	12	0	0	2010
			U	23	30	126	6454	3691	496	1946	2445	4

Aumentando il valore di **numClusters** la precisione del clustering aumenta di conseguenza. Con 6 centroidi gli spermatozoi si dividono in due cluster caratterizzati da bassissima entropia e nel complesso con un alto valore di recall. Aumentando ulteriormente il numero di cluster diminuisce l'entropia ma la purezza dei cluster non varia significativamente.

3.1.2 Considerazioni sulla terna di *feature* FSLog-SSLog-PI

Utilizzare le tre *feature* FSLog-SSLog-PI cercando un numero di cluster superiore a 2 permette di individuare cluster di spermatozoi con bassa entropia, un alto grado di purezza e nel complesso un recall elevato.

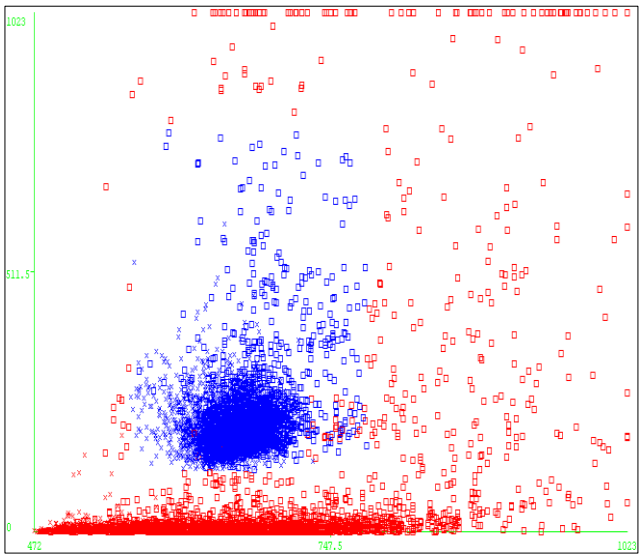


Figura 20: Classificazione dell'esperto del dominio (bi-plot FSL-PI). In blu gli spermatozoi, in rosso le altre cellule

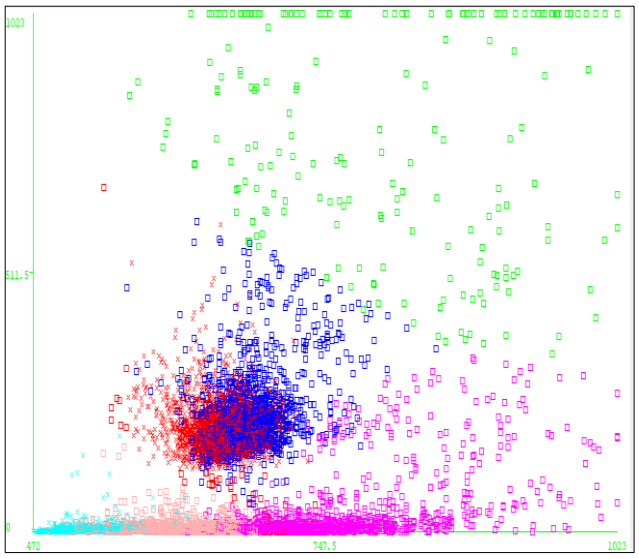


Figura 21: Risultato del clustering (bi-plot FSL-PI) con 6 cluster

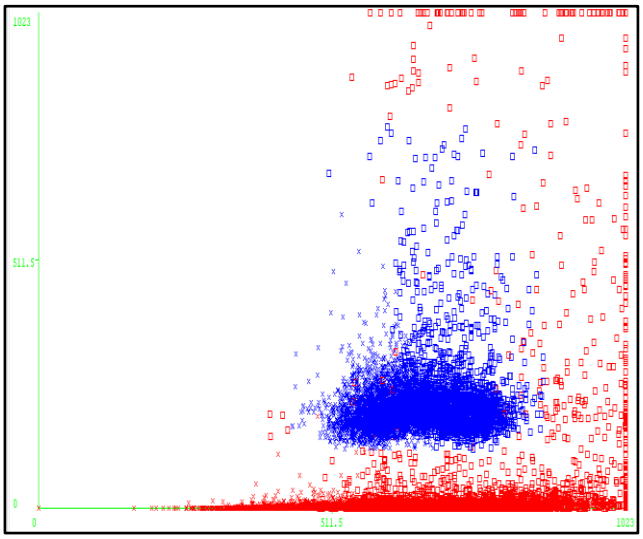


Figura 22: Classificazione dell'esperto del dominio (bi-plot SSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

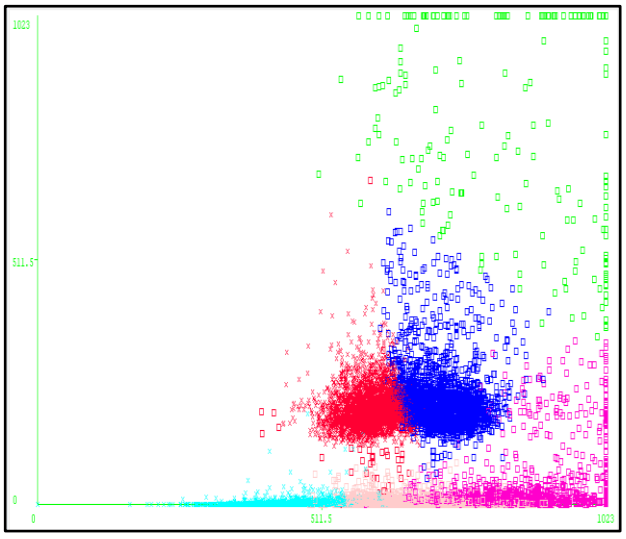


Figura 23: Risultato del clustering (bi-plot SSL – PI)con 6 cluster

In generale il problema che si presenta con K-means è che per avere buoni risultati bisogna impostare il numero di cluster superiore a 2. Non si otterrà quindi in output dall'algoritmo un solo cluster di spermatozoi e un solo cluster unknown, ma si otterranno una serie di

cluster, molto omogenei, che in un passaggio successivo dovranno essere in qualche modo etichettati in cluster di spermatozoi e cluster di cellule che non sono spermatozoi.

3.1.3 Identificazione del miglior valore dei parametri

Basandosi sui risultati ottenuti con il data set 68i+ Z0019566 si sono voluti individuare i valori per i parametri caratteristici di K-Means che permettessero il migliore riconoscimento degli spermatozoi. Nel cercare la configurazione migliore si è tenuto conto non solo del numero degli spermatozoi individuati ma anche dei falsi positivi.

Dal risultato dei test svolti precedentemente, il valore del **seed** non ha influito sul risultato finale. Per questo motivo si è deciso di lasciare settato il **seed** al valore di default 10.
La funzione di distanza e il numero di cluster hanno rappresentato i parametri da ottimizzare.

# cluster	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
Euclidean Distance													
4	0,050567	0,992561	S	5832	1	56	2						
			U	98	5016	882	9219						
5	0,051458	0,99114	S	64	5772	53	0	2					
			U	164	68	1448	8146	5389					
Manhattan Distance													
6	0,052676	0,992798	S	0	28	3707	2	2147	7				
			U	5356	1318	59	3533	56	4893				
10	0,033573	0,995594	S	0	2	2389	1497	1967	0	22	11	3	0
			U	3314	2896	26	19	10	1990	859	156	2024	3921

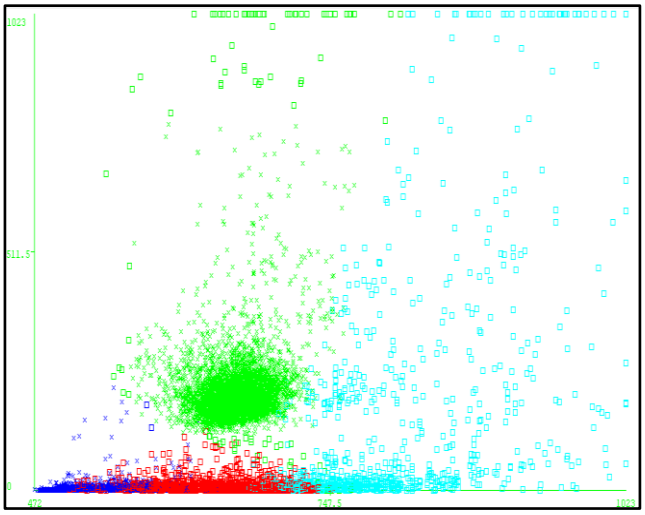


Figura 24: #Cluster = 4; Euclidean Distance

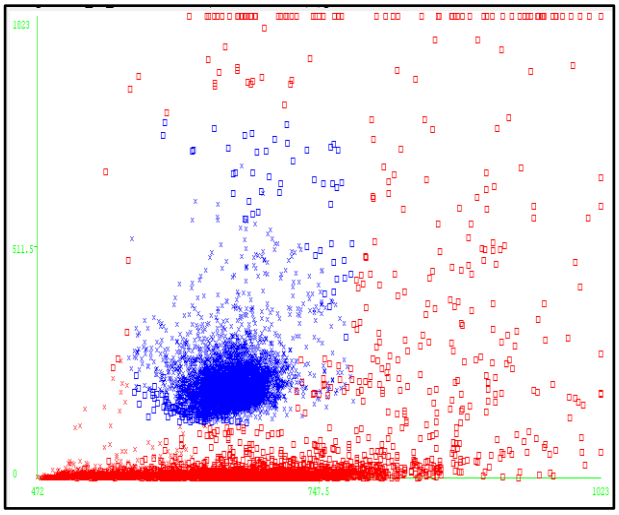


Figura 25: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

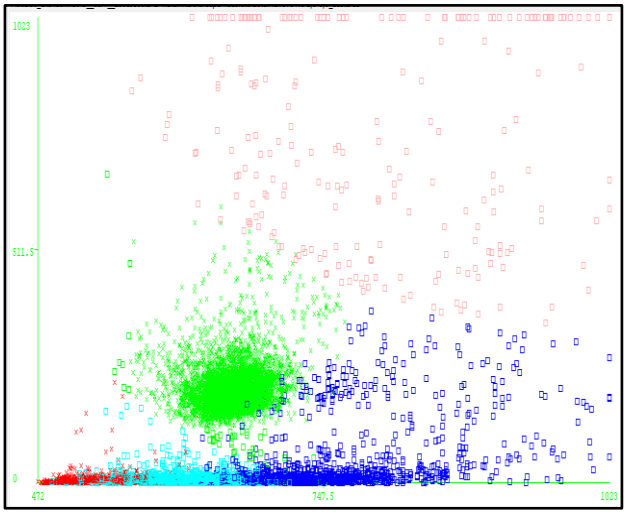


Figura 26: #Cluster = 5; Euclidean Distance

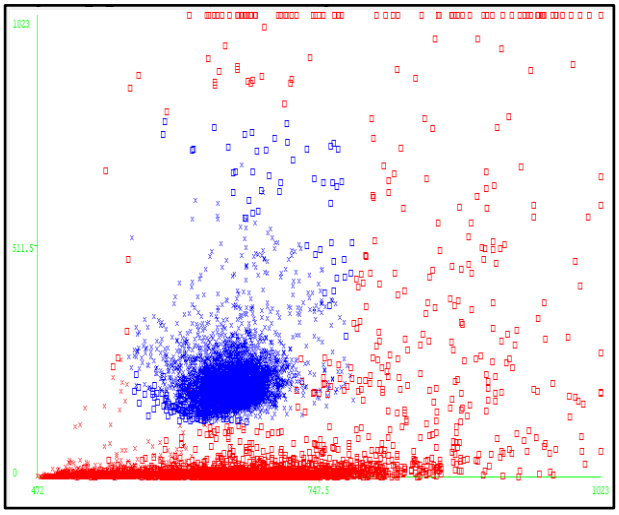


Figura 27: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

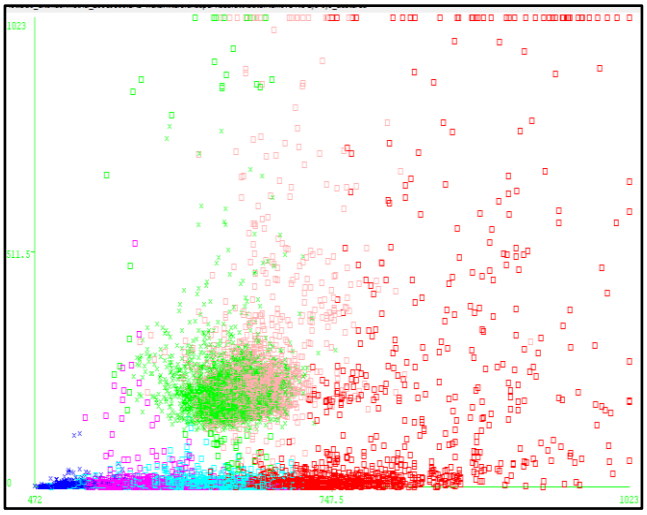


Figura 28: #Cluster = 6; Manhattan Distance

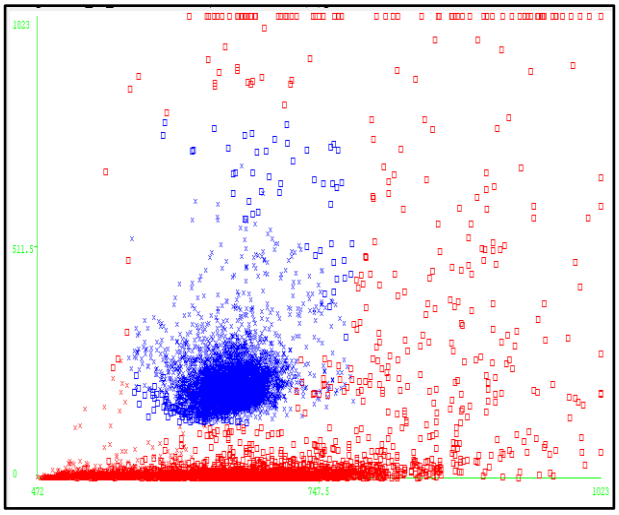


Figura 29: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

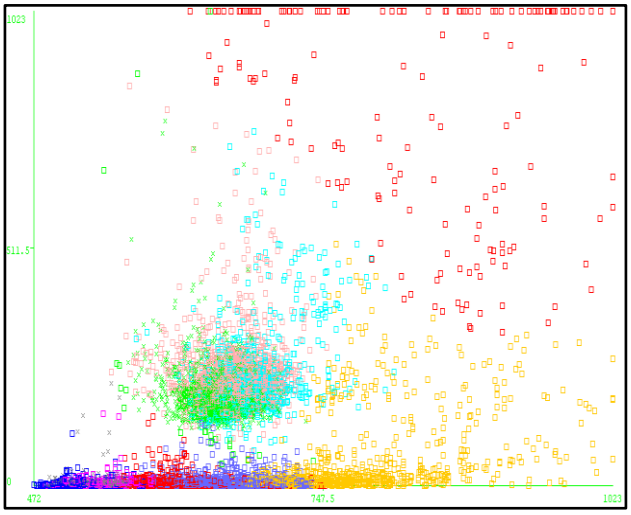


Figura 30: #Cluster = 10; Manhattan Distance

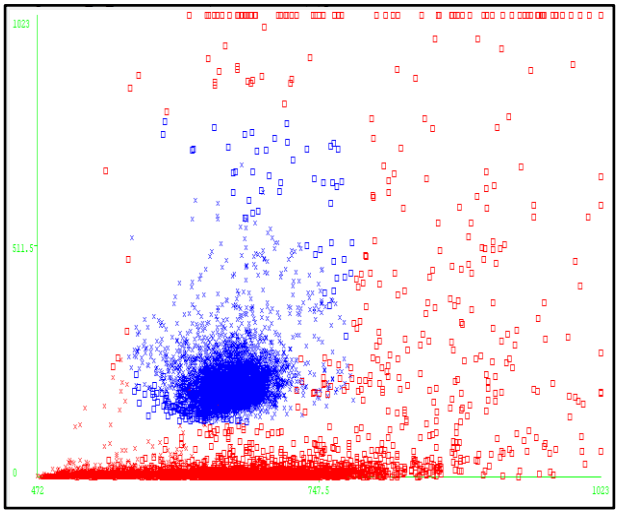


Figura 31: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

Con la funzione di distanza di Manhattan si sono ottenuti i risultati migliori nel caso di spermatozoi suddivisi in più di un cluster. La distanza Euclidea si è comportata meglio nei casi in cui gli spermatozoi venivano inclusi in un unico cluster.

Per quanto riguarda il numero di cluster i risultati migliori si sono ottenuti con un valore superiore a 4. Con 4 e 5 centroidi gli spermatozoi vengono inseriti in un unico cluster. Aumentando i numeri dei centroidi gli spermatozoi si suddividono in 2 o 3 cluster.

3.1.4 Validazione dei parametri ottimali con altri data set

Identificati i valori ottimali per i parametri di K-Means, l'algoritmo è stato testato su altri due campioni per verificare le sue capacità di generalizzazione.

Per poter eseguire una valutazione il più possibile attendibile sono stati scelti due data set acquisiti con le stesse impostazioni del citofluorimetro del campione utilizzato nella fase esplorativa. I campioni utilizzati sono stati: 68i- Z0019565 e 63i- Z0019561.

3.1.4.1 Campione Z0019565

# cluster	Entropia Clustering	Purezza Clustering	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	C2	C3	C4
				S	U	S	U						
Euclidean Distance													
4	0,14814	0,943568	0,14482409	0,979	0,021	0,858	0,010	S	0	1467	8936	6	
								U	10359	1445	188	6643	
5	0,1036	0,977861	0,11524295	0,984	0,016	0,953	0,008	S	6	112	9922	369	0
								U	7112	197	156	2030	9140

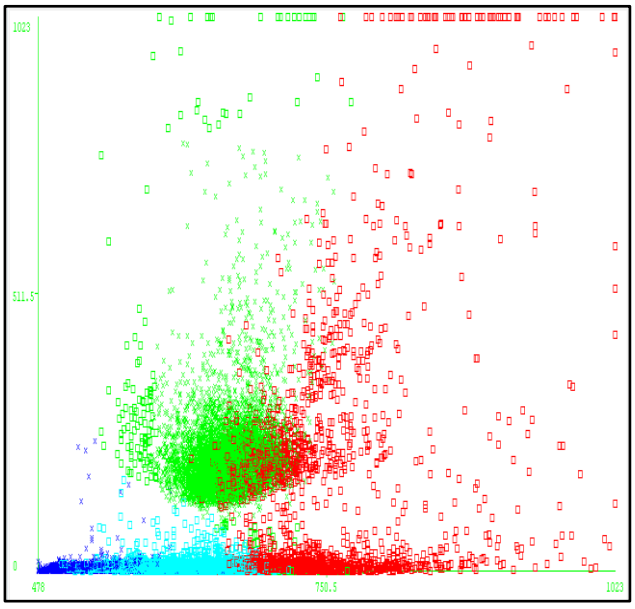


Figura 32: #Cluster = 4; Euclidean Distance

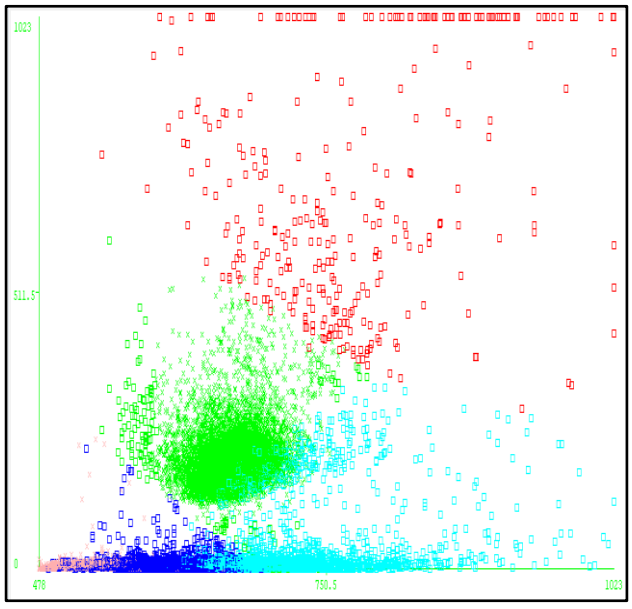


Figura 33: #Cluster = 5; Euclidean Distance

3.1.4.2 Campione Z0019561

# cluster	Entropia Clustering	Purezza Clustering	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	C2	C3	C4
				S	U	S	U						
Euclidean Distance													
4	0,310817	0,912836	0,20738558	0,967	0,033	0,714	0,034	S	2951	38	7502	2	
								U	1262	523	253	5309	
5	0,199151	0,947365	0,04468950	0,995	0,005	0,715	0,005	S	13	26	7506	1	2947
								U	2145	497	37	3806	862

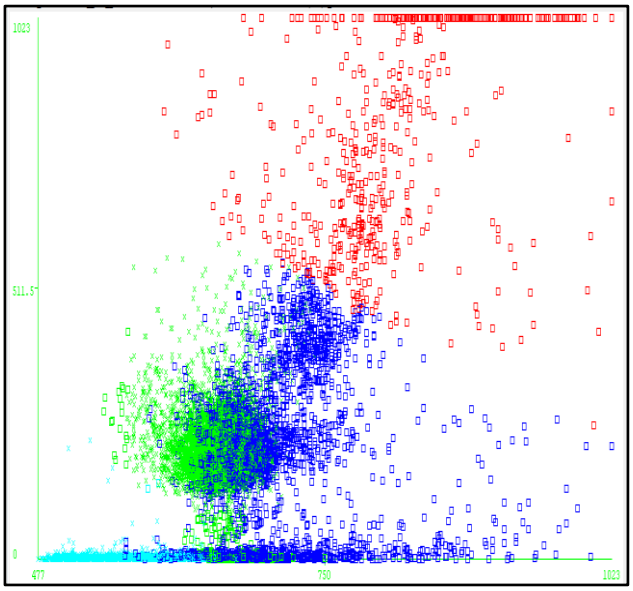


Figura 34: #Cluster = 4; Euclidean Distance

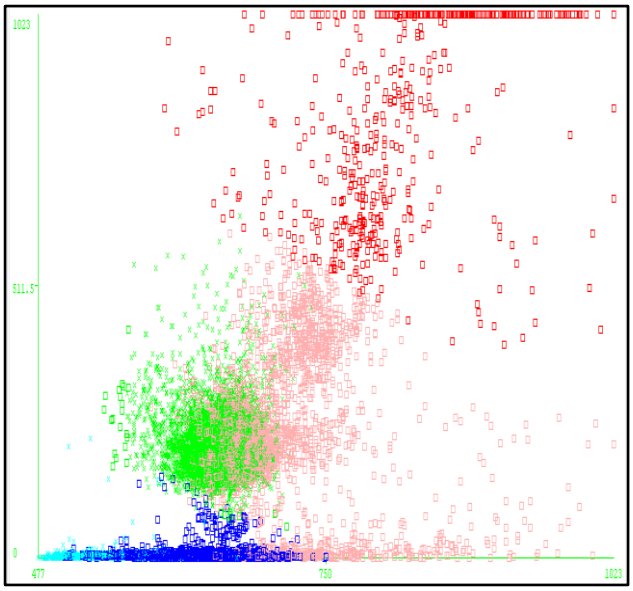


Figura 35: #Cluster = 5; Euclidean Distance

3.1.4.3 Considerazioni

L'algoritmo K-Means applicato ad altri data set non ha confermato le buone prestazioni ottenute durante l'analisi esplorativa. Gli spermatozoi tendono a suddividersi in più cluster con bassi valori di purezza e recall.

3.2 Expectation Maximisation

Expectation Maximisation (EM) è una tecnica di clustering basata su un modello probabilistico *[11]*. In un modello probabilistico l'obiettivo è individuare uno schema di clustering che mappa nella maniera migliore un certo modello parametrico.
Il modello statistico utilizzato è chiamato *finite mixtures*. Questo modello prevede di rappresentare *k* cluster con *k* distribuzioni di probabilità le quali determinano i valori degli attributi delle istanze che fanno parte del cluster. In pratica ogni distribuzione determina la probabilità di un' istanza del data set di avere un certo valore nei suoi attributi se facesse parte di quel cluster.

Il modello statistico più semplice prevede un insieme di *k* Gaussian mixture nel quale ognuno dei *k* cluster ha un valore di media e varianza differente. L'obiettivo è determinare a quale cluster ogni istanza appartiene tenendo presente che la probabilità di appartenenza di una istanza ad un cluster non è equamente distribuita.

La funzione di densità di probabilità del modello finite Gaussian mixture si presenta come una "catena montuosa" con un picco per ogni componente*[11]*; in Figura 36 ne è mostrato un esempio. Il modello ipotizza la presenza di due cluster A e B con i relativi valori di valor medio μ_A, μ_B e deviazione standard σ_A, σ_B. I campioni sono presi da queste distribuzioni attingendo al cluster A con probabilità p_A e al cluster B con probabilità p_B (dove $p_A + p_B = 1$). Se ora si suppone di avere solo il data set, senza sapere a che cluster appartiene ogni istanza, il problema di un modello finite mixture consiste nel determinare i cinque parametri μ_A, μ_B, σ_A, σ_B e p_A (il valore di p_B è ricavabile direttamente da p_A).

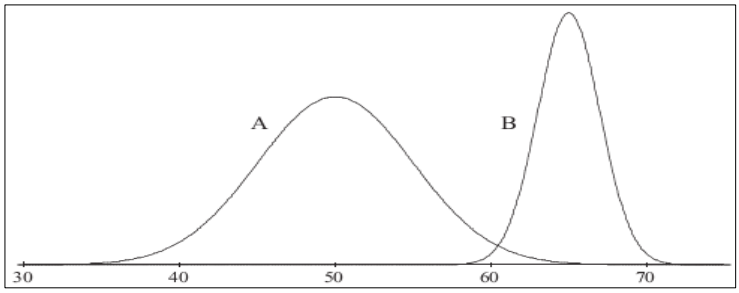


Figura 36: Esempio di finite Gaussian mixture

Estendendo il problema a *k* cluster bisognerà individuare i valor medi, varianze e probabilità di appartenenza di un'istanza ad un cluster per tutti i *k* cluster.

L'algoritmo EM, come K-Means, utilizza un approccio iterativo che segue questi passi:

- Fase 1: si scelgono in modo random i valori μ_K, σ_K, e p_K
- Fase 2: si utilizza la funzione di densità di probabilità per distribuzioni gaussiane per calcolare le probabilità di appartenenza a ogni cluster di ogni istanza (expectation).
- Fase 3: si utilizzano le probabilità stimate per calcolare nuovamente i parametri μ_K, σ_K, e p_K (maximization of the likelihood[17]), e si ripetono le fasi 2 e 3 fino a quando la misura di qualità dei cluster, la likelihood, non migliora di una quantità significativa.

L'algoritmo Expectation Maximisation è stato applicato al data set variando i parametri caratteristici. L'implementazione utilizzata prevede un **seed** che permette di inizializzare in modo random i parametri iniziali di μ_K, σ_K, e p_K. Non sono state però riportate le prove effettuate poiché variandone il valore lo schema di clustering risultante non ha mai presentato significative differenze.
Anche per il valore di **maxIterations** non sono state riportate le diverse prove poiché a posteriori si è visto che il numero di iterazioni massimo non ha mai raggiunto il valore limite.
Sono stati effettuati esperimenti variando il valore di **minStdDev**[18] e **numCluster** . Per la precisione, la maggior parte delle volte, la scelta del numero di cluster da individuare è stata determinata automaticamente dall'algoritmo attraverso la cross-validazione secondo questa procedura:

1. il numero di cluster è posto a 1.
2. il training set è splittato in modo random in 10 partizioni.
3. EM è eseguito 10 volte usando iterativamente 1 partizione come test e le altre come training.
4. la misura di loglikelihood è la media dei 10 risultati.
5. se la misura di loglikelihood è aumentata allora si aumenta di 1 il numero di cluster e si riparte dallo step 2.

[17] Likelihood è una misura di verosimiglianza dei dati rispetto al modello determinato dal clustering.
[18] Valore di deviazione standard minima ammissibile per un cluster

3.2.1 *Feature* FSLog-SSLog-PI

3.2.1.1 Deviazione standard minima

minStdDev	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
1.0E-6 – 0.4	0,048609	0,985123	S	35	0	5389	467	0	0	0
			U	439	3075	16	263	3583	5446	2393
0.5 – 0.6	0,066357	0,980243	S	0	0	5498	393			
			U	5036	9192	24	963			
0.7 – 1.2	0,353231	0,879323	S	0	5891					
			U	12668	2547					
1.3 – 1.4	0,061029	0,985312	S	36	0	5617	31	0	207	
			U	369	1883	64	1627	11093	179	
2.0	0,060299	0,985691	S	39	0	5642	28	0	182	
			U	411	2065	65	2188	10316	170	
5.0	0,077178	0,986212	S	48	37	5626	180			
			U	5007	9465	26	717			
10	0,063868	0,983275	S	5565	0	260	66	0		
			U	27	1109	302	7141	6636		
20	0,082249	0,980148	S	16	190	5479	9	12	185	
			U	511	2199	7	5669	6639	190	
50	0,057477	0,990382	S	92	24	20	5719	27	9	
			U	184	5054	463	31	6833	2650	
60	0,07823	0,986544	S	118	24	10	5647	40	52	
			U	186	5393	544	40	6465	2587	
70	0,06252	0,989434	S	15	67	5709	0	100		
			U	1141	8968	41	4853	212		
80	0,082383	0,98896	S	60	51	5703	77			
			U	3579	577	45	11014			
90	0,063064	0,989624	S	0	55	5755	0	1	80	
			U	1811	7184	83	440	5533	164	
100	0,208293	0,950109	S	0	0	5807	17	0	67	
			U	954	863	969	11927	341	161	
110	0,295723	0,938122	S	67	357	0	0	5444	0	23
			U	153	12641	531	360	872	648	10
120	0,70742	0,802852	S	2905	74	2912				
			U	13784	249	1182				
130	0,682286	0,81465	S	0	2820	3004	0	0	67	
			U	0	13763	1025	274	0	153	

Lasciando decidere all'algoritmo il numero migliore di cluster, se la **minStdDev** si mantiene entro il valore di 100, gli spermatozoi tendono a dividersi tra un cluster molto rappresentativo con un alto valore di recall ed i rimanenti a distribuirsi in tanti altri cluster poco rappresentativi.

Da notare che nell'esperimento in cui **minStdDev** era compreso in [0.7...1.2] si sono ottenuti due cluster solo a causa del procedimento seguito dall'algoritmo per il quale se il valore di loglikelihood non aumenta non si prova con un numero maggiore di cluster. Impostando manualmente un numero di cluster maggiore di 2 si ottengono risultati in linea con quanto è stato descritto in precedenza e con 8 cluster il valore di loglikelihood è molto simile a quello ottenuto dall'algoritmo con 2 soli centroidi.

Impostando un valore di **minStdDev** oltre 100, la qualità del clustering degenera progressivamente e gli spermatozoi tendono a dividersi in più cluster sempre più caratterizzati da alta entropia.

3.2.2 Considerazioni sulla terna di *feature* FSLog-SSLog-PI

L'utilizzo della terna FSLog-SSLog-PI si è dimostrato efficace nel riuscire a individuare uno schema di clustering nel quale gli spermatozoi sono ben distinti dalle altre cellule. I migliori risultati si sono ottenuti nei casi in cui il numero di cluster era compreso tra [4..5] e il valore di **minStdDev** compreso tra [70..90].
Con valori di **minStdDev** minori di 70 gli spermatozoi tendono a distribuirsi anche in altri cluster, mentre con un valore maggiori di 90 il cluster degli spermatozoi diventa progressivamente più impuro.

Tabella 1: weka.clusterers.EM -I 400 -N -1 -M 90.0 -S 100

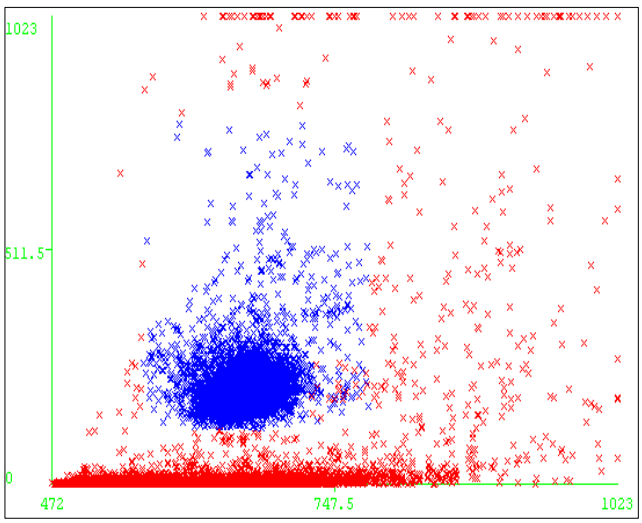


Figura 37: Classificazione dell'esperto del dominio (FSL-PI). In blu gli spermatozoi, in rosso le altre cellule

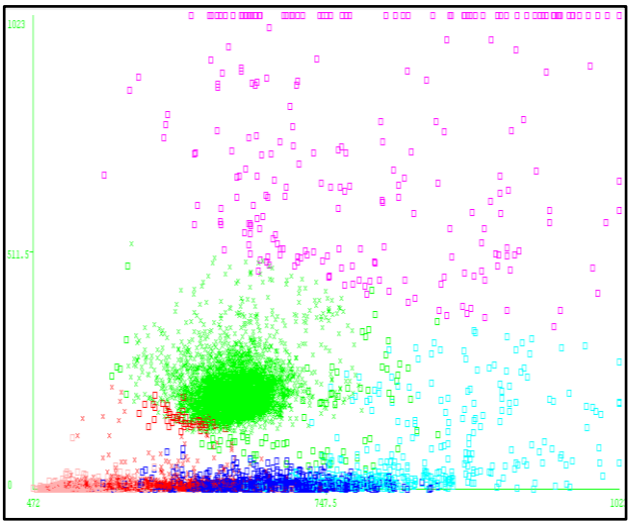


Figura 38: Risultato del clustering (FSL-PI) con 6 cluster

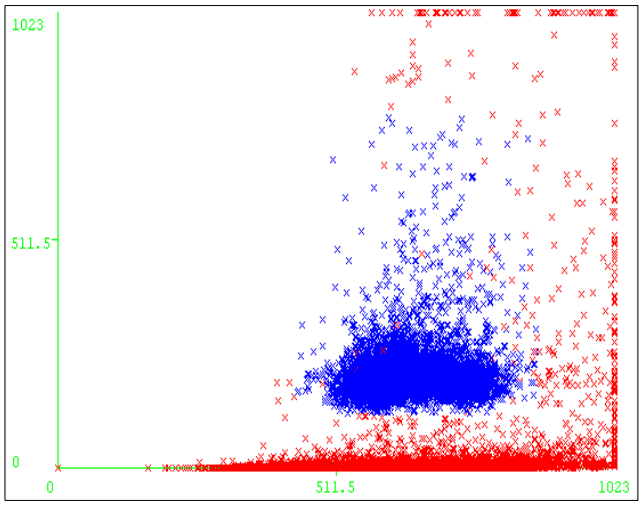


Figura 39: Classificazione dell'esperto del dominio (SSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

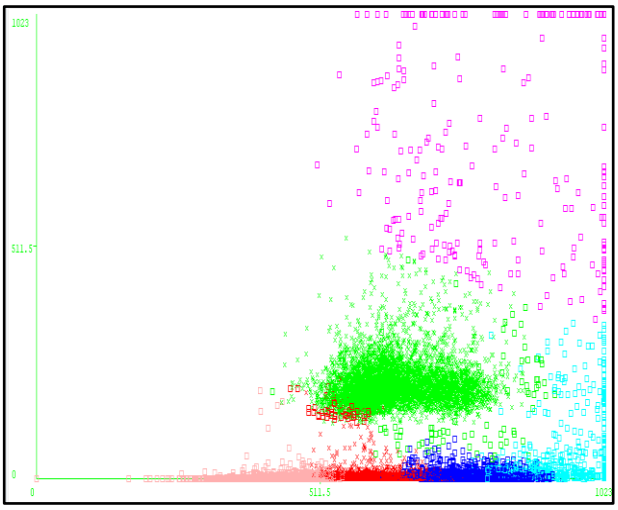


Figura 40: Risultato del clustering (SSL-PI)

3.2.3 Identificazione del miglior valore dei parametri

Basandosi sui risultati ottenuti con il data set Z0019566 si sono voluti individuare i migliori valori per i parametri caratteristici di Expectation Maximisation. Nel cercare la configurazione migliore si è tenuto conto non solo del numero degli spermatozoi individuati ma anche dei falsi positivi.

Dal risultato dei test svolti precedentemente, i valori di **seed e maxIterations** non hanno influito sul risultato finale. Per questo motivo si è deciso di lasciare settato il **seed** al valore di default 100 e **maxIterations** al valore 400. I parametri da ottimizzare sono stati il numero di cluster e il valore di deviazione standard.

minStdDev	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
70	0,06252	0,989434	S	15	67	5709	0	100		
			U	1141	8968	41	4853	212		
80	0,082383	0,98896	S	60	51	5703	77			
			U	3579	577	45	11014			
80	0,051505	0,990714	S	0	2	0	76	5811	0	2
			U	355	9269	3	156	116	2187	3129
90	0,063064	0,989624	S	0	55	5755	0	1	80	
			U	1811	7184	83	440	5533	164	

Il più alto valore di recall con un numero di falsi positivi limitato si è ottenuto impostando un numero di cluster pari a 7 e una deviazione standard minima pari a 80.

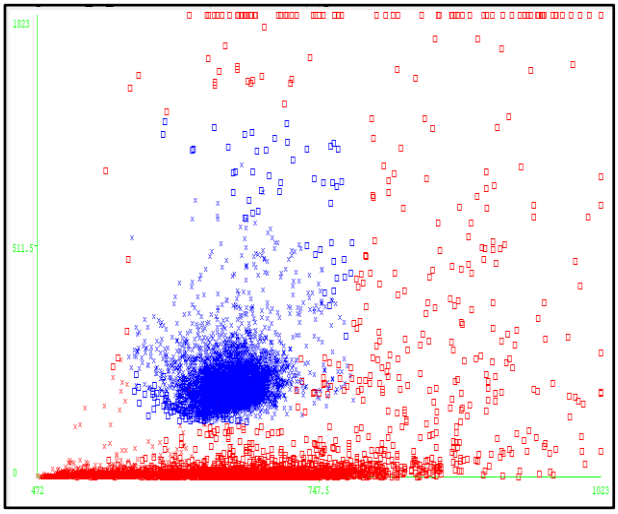


Figura 41: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

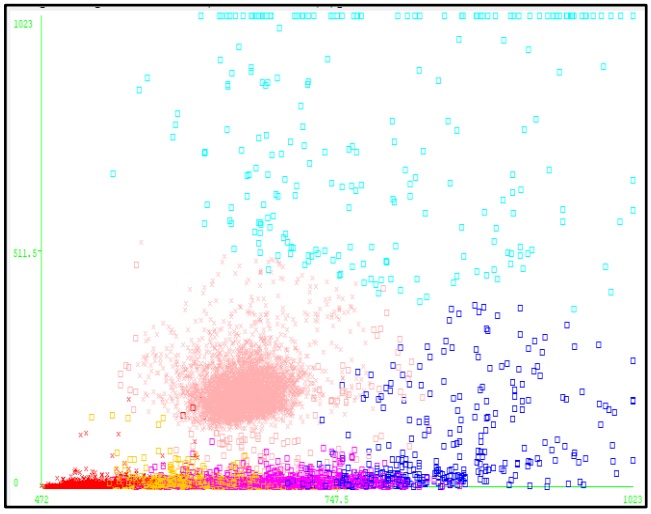


Figura 42: Risultato del clustering (FSL-PI)

3.2.4 Validazione dei parametri ottimali con altri data set

Identificati i valori ottimali per i parametri di Expectation Maximisation, l'algoritmo è stato testato su altri due campioni per verificare le sue capacità di generalizzazione.

Per poter eseguire una valutazione il più possibile attendibile sono stati scelti due data set acquisiti con le stesse impostazioni del citofluorimetro del campione utilizzato nella fase esplorativa. I campioni utilizzati sono stati: 68i- Z0019565 e 63i- Z0019561.

3.2.4.1 Campione Z0019565

min StdDev	Entropia Clustering	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	C2	C3	C4	C5	C6
			S	U	S	U								
80	0,068922	0,13082496	0,981	0,019	0,946	0,010	S	415	0	0	0	23	9855	116
							U	31	5333	492	3226	9179	182	192

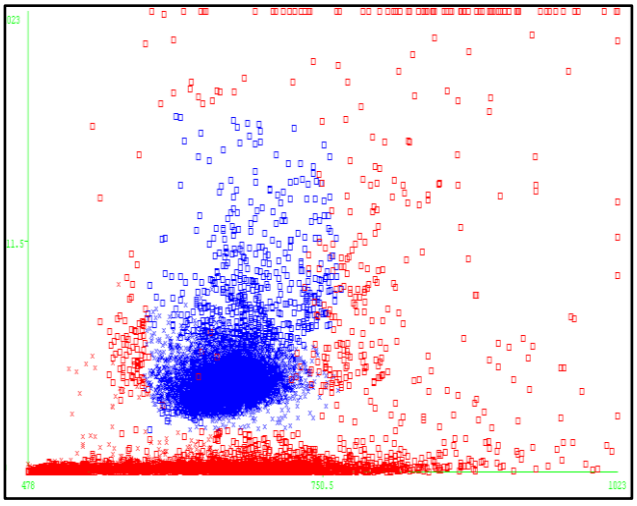


Figura 43: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

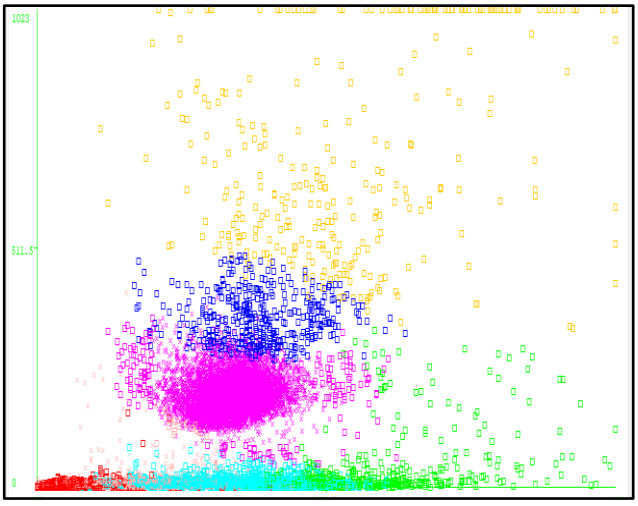


Figura 44: Risultato del clustering (FSL-PI)

3.2.4.2 Campione Z0019561

min StdDev	Entropia Clustering	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	C2	C3	C4	C5	C6
			S	U	S	U								
80	0,292377	0,38805698	0,923	0,077	0,947	0,111	S	9940	217	1	17	0	0	318
							U	818	474	3679	1765	277	262	72

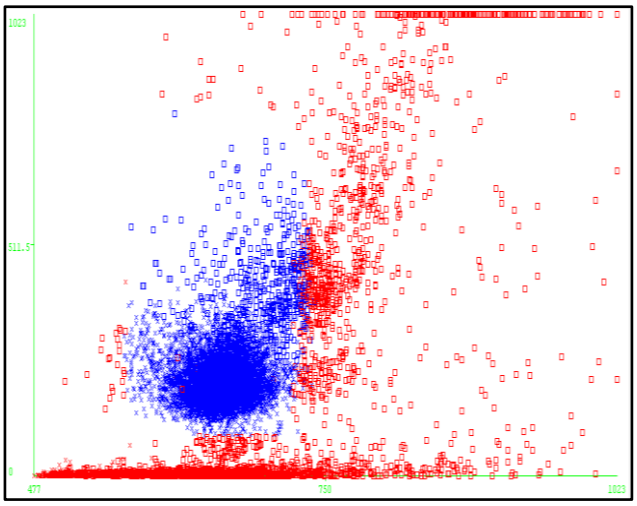


Figura 45: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

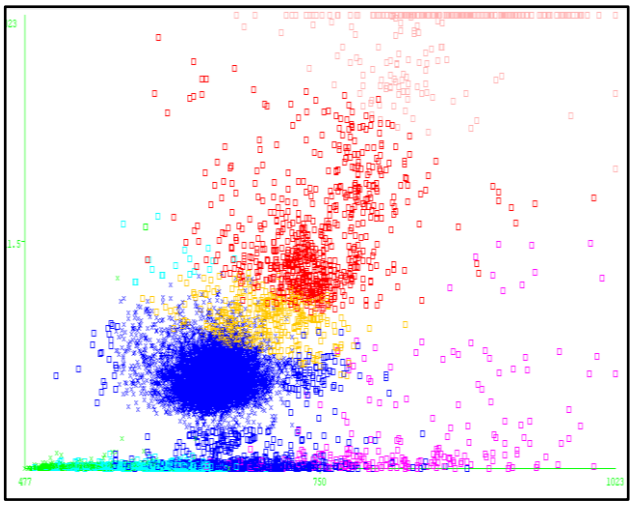


Figura 46: Risultato del clustering (FSL-PI)

3.2.4.3 Considerazioni

L'algoritmo EM applicato ai due campioni di test non è stato efficace come nel test set: gli spermatozoi si sono suddivisi tra più cluster di cui il più rappresentativo caratterizzato da un valore di recall medio del 94%, però sia questo cluster che gli altri contenenti gli spermatozoi includono un numero elevato di falsi positivi.

3.3 DBScan

DBScan è una tecnica di clustering basata sul concetto di densità *[9]*. Obiettivo del clustering è individuare nello spazio delle regioni ad alta densità separate le une dalle altre da regioni a bassa densità. DBScan valuta la densità con un approccio center-based, cioè conta il numero di punti all'interno di un determinato raggio, epsilon, dal punto del data set preso in esame. Con questo concetto la densità di ogni punto dipenderà dal valore del raggio utilizzato e questo determinerà il numero di cluster identificabili nel data set.
L'approccio utilizzato in DBScan permette di classificare i punti in tre gruppi:

- Core point: punti appartenenti a regioni dense. Un punto è considerato core point se il numero di punti all'interno del suo vicinato determinato da una funzione di distanza e dal valore epsilon supera un valore soglia, minPoints.
- Border point: punti che si trovano ai bordi di regioni dense. Un border point non è un core point ma è un punto che si trova nel vicinato di un core point.
- Noise point: punti che si trovono in regioni a bassa densità. I noise point non sono core point né border point.

L'algoritmo segue queste fasi *[9]*:

- Fase 1: tutte le istanze vengono classificate come core, border o noise point secondo le definizioni precedenti
- Fase 2: tutti i noise point vengono scartati
- Fase 3: vengono idealmente collegati tra loro i core point che si trovano entro una distanza epsilon l'uno dall'altro.
- Fase 4: ogni gruppo di core point connessi è identificato come cluster
- Fase 5: ogni border point viene associato ad uno dei cluster che contiene un suo core point

Prima di applicare l'algoritmo di clustering gli attributi FSLog, SSLog e PI del data set sono stati normalizzati nell'intervallo [0..1] al fine di comprendere meglio i risultati al variare dei due parametri caratteristici: **ε** e **MinPoints.** L'operazione di normalizzazione si è resa necessaria in quanto il software di data mining Weka nell'applicare l'algoritmo DBScan, seppure visualizzi il risultato con il valore degli attributi degli oggetti originari, normalizza automaticamente ogni singolo attributo nel range [0..1] e successivamente utilizza il valore di **ε** e **MinPoints** per trovare i cluster.

L'algoritmo DBScan è stato applicato al data set variando i parametri caratteristici **ε** e **MinPoints**.

3.3.1 *Feature* FSLog-SSLog-PI

3.3.1.1 Epsilon = 0.01

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
150	19623	0	1	S	0	0					
				U	741	742					
50	12977	0	1	S	0						
				U	8129						
20	8965	- (16 cluster)	- (16 cluster)	S	0	-	783	-	-	-	-
				U	10854	-	0	-	-	-	-
5	3288	- (61 cluster)	- (61 cluster)	S	3949	0	-	-	-	-	
				U	0	13383	-	-	-	-	-

Il valore di **ε** così basso non permette di distinguere nessun cluster di interesse. In generale il numero di cellule considerate outlier è molto elevato.

Con un numero ridotto di **minPoints** si creano una moltitudine di micro-cluster.

3.3.1.2 Epsilon = 0.02

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6	C7
300	14788	0	1	S	0							
				U	6318							
150	10624	0	1	S	1064	0						
				U	0	9418						
70	6321	0	1	S	2993	0						
				U	0	11792						
30	3046	0	1	S	4614	0	0					
				U	0	13388	58					
15	2006	0,001333	0,999895	S	5001	0	0					
				U	2	14074	23					
7	1390	0,001299	0,999899	S	5281	0	0	0	7	0	0	0
				U	2	14368	18	17	0	9	7	7
3	874	- (56cluster)	- (56cluster)	S	5434	0	-	-	-	-	-	-
				U	2	14559	-	-	-	-	-	-

Con un tale valore di **ε** si incominciano a distinguere due cluster, quello degli spermatozoi e quello unknown, a patto che il valore di **minPoints** non sia troppo basso. Al di sotto del valore 15 in aggiunta ai due cluster principali si riscontrano una moltitudine di micro-cluster che non hanno alcun significato apparente.

3.3.1.3 Epsilon = 0.03

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
150 - 50	4391	0	1	S	4474	0					
				U	0	12241					
20	1212	0,004363	0,999598	S	5426	0					
				U	8	14460					
10	980	0,004309	0,999604	S	5535	0	0	0			
				U	8	14629	16	10			
5	617	- (21 cluster)	- (21 cluster)	S	5621	0	-	-	-	-	-
				U	14	14723	-	-	-	-	-
2	309	- (109 cluster)	- (109 cluster)	S	5651	0	-	-	-	-	-
				U	20	14736	-	-	-	-	-

Con un tale valore di **ε** si distinguono sempre due cluster, quello degli spermatozoi e quello unknown, a partire da valori di **minPoints** elevati intorno a 150. Diminuendo il valore di **minPoints** continua ad essere ben rappresentato il cluster degli spermatozoi e allo stesso tempo si crea una moltitudine di micro-cluster che non hanno alcun significato apparente.

3.3.1.4 Epsilon = 0.04

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6	C7
300	4054	0,002134	0,999824	S	4953	0						
				U	3	12096						
150	2450	0,005625	0,999464	S	5302	0						
				U	10	13344						
30 - 15	949	0,007425	0,999256	S	5554	0						
				U	15	14588						
7	472	0,847505	0,724338	S	5688	25	12	0	0	7	0	7
				U	14859	0	0	20	9	0	7	0

Con un tale valore di **ε** è possibile individuare immediatamente il cluster degli spermatozoi e quello unknown soprattutto con valori di **minPoints** elevati.

Diminuendo il valore di **minPoints** sotto la decina il cluster degli spermatozoi e quello degli unknown si uniscono.

3.3.1.5 Epsilon = 0.05

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
300	1813	0,009044	0,999067	S	5476	0					
				U	18	13799					
150	1452	0,010132	0,998932	S	5518	0					
				U	21	14115					
30	687	0,012858	0,99858	S	5653	0					
				U	29	14737					
15	490	0,849635	0,723273	S	5705	27	15	0			
				U	14853	0	0	16			

Similmente al caso con **ε = 0.04** è possibile individuare immediatamente il cluster degli spermatozoi e quello unknown con valori di **minPoints** elevati.
Diminuendo **minPoints** sotto il valore 30 il cluster degli spermatozoi e quello degli unknown si uniscono.

3.3.1.6 Epsilon = 0.08

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
150 - 800	660	0,854858	0,720434	S	5716						
				U	14730						
5	38	0,852554	0,720714	S	5882	0	2				
				U	15096	85	3				

Qualunque sia il valore di **minPoints** è stato impossibile riuscire ad isolare gli spermatozoi. Con valori di **minPoints** intorno a 5 si sono però individuati cluster di cellule che potrebbero essere interessanti per l'esperto del dominio nello studio di particolari cellule.

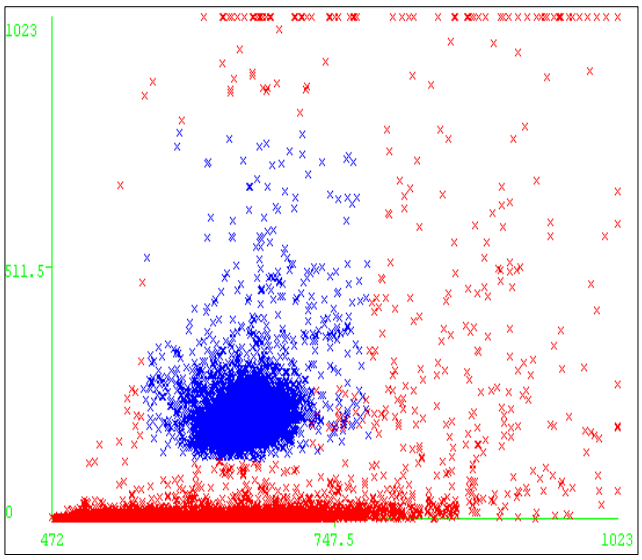


Figura 47: Classificazione dell'esperto del dominio (FSL-PI). In blu gli spermatozoi, in rosso le altre cellule

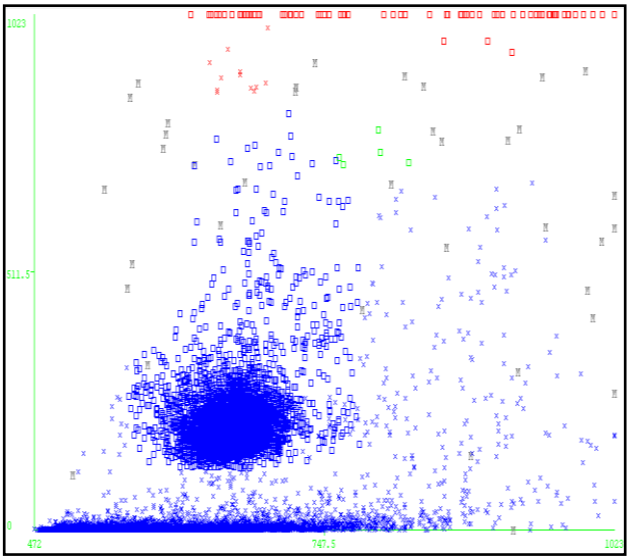


Figura 48: Cluster diversi sono colorati diversamente (epsilon = 0.08, minPoints = 5).

3.3.1.7 Epsilon = 0.1

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
>= 50	299	0,856564	0,719181	S	5843						
				U	14964						
10	35	0,852444	0,720754	S	5884	0	0				
				U	15099	45	43				

Per qualunque valore di **minPoints** non è stato possibile riuscire ad isolare gli spermatozoi. Si sono però individuati cluster di cellule che potrebbero essere interessanti per l'esperto del dominio.

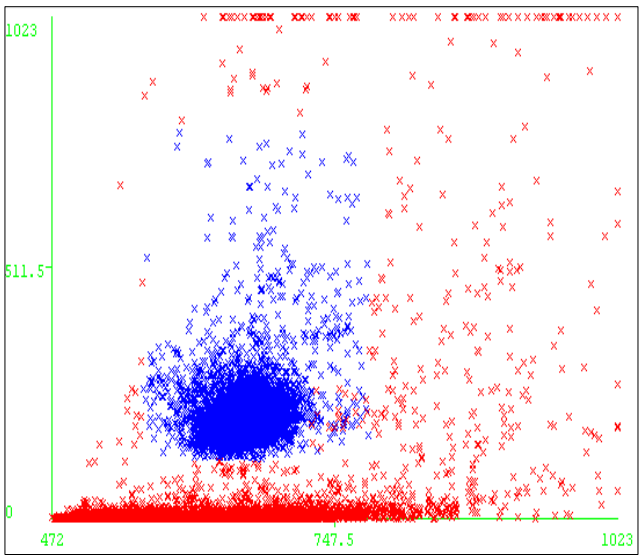


Figura 49: Classificazione dell'esperto del dominio (FSL-PI). In blu gli spermatozoi, in rosso le altre cellule

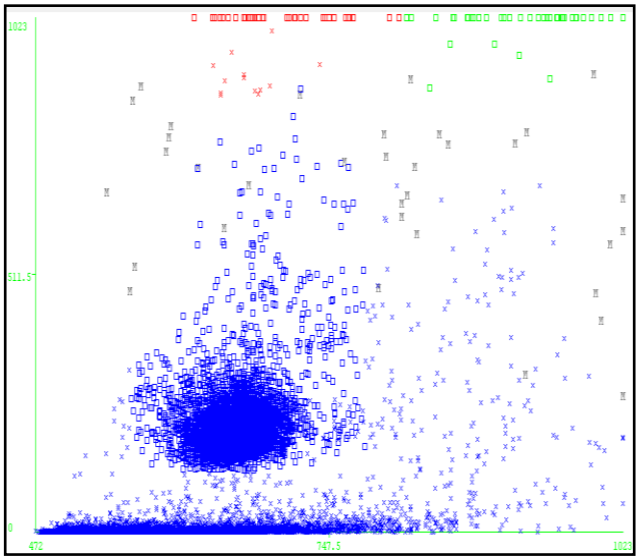


Figura 50: Cluster diversi sono colorati diversamente (epsilon = 0.1, minPoints = 10).

3.3.1.8 Epsilon = 0.5

minPoints	#oggetti noise	Entropia Clustering	Purezza Clustering		C0	C1	C2	C3	C4	C5	C6
>= 600	11	0,854441	0,72074	S	5891						
				U	15204						
10	0	0,854242	0,720885	S	5891						
				U	15215						

Il valore di **ε** particolarmente elevato non ha permesso di separare gli spermatozoi dalle altre cellule. Tutte le istanze del campione sono confluite in un unico cluster.

3.3.2 Considerazioni sulla terna di *feature* FSLog-SSLog-PI

L'utilizzo della terna FSLog-SSLog-PI si è dimostrato efficace nel riuscire a individuare uno schema di clustering nel quale gli spermatozoi sono ben distinti dalle altre cellule

Tabella 2: clustering ottenuto con le seguenti impostazioni
weka.clusterers.DBScan -E 0.04 -M 30 -I weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase -D weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclidianDataObject

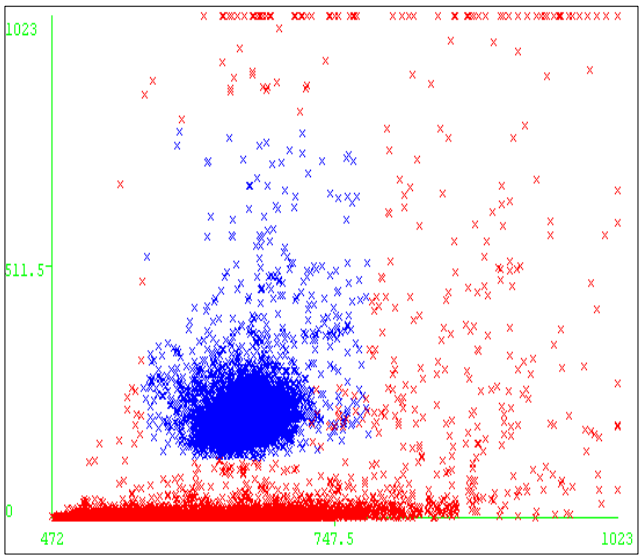


Figura 51: Classificazione dell'esperto del dominio (FSL-PI). In blu gli spermatozoi, in rosso le altre cellule

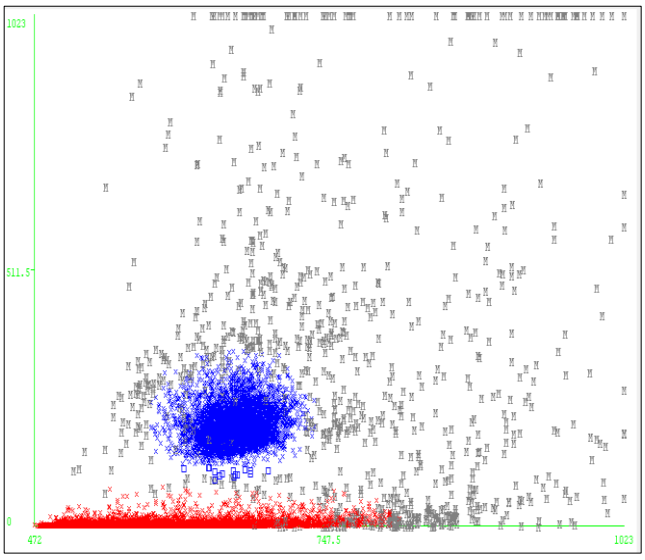


Figura 52: Risultato del clustering (FSL-PI)

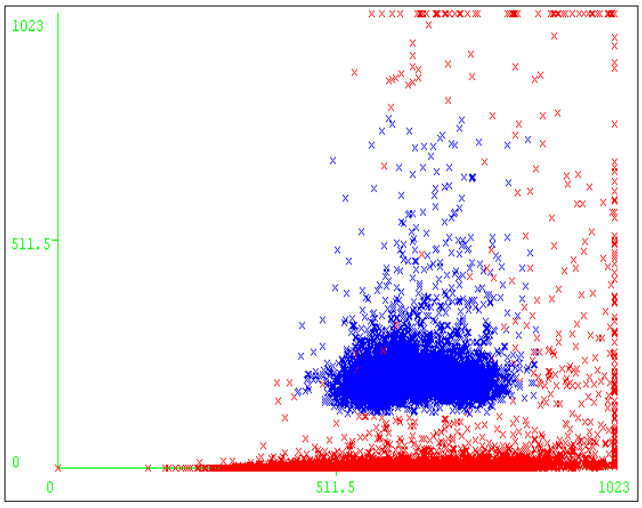


Figura 53: Classificazione dell'esperto del dominio (SSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

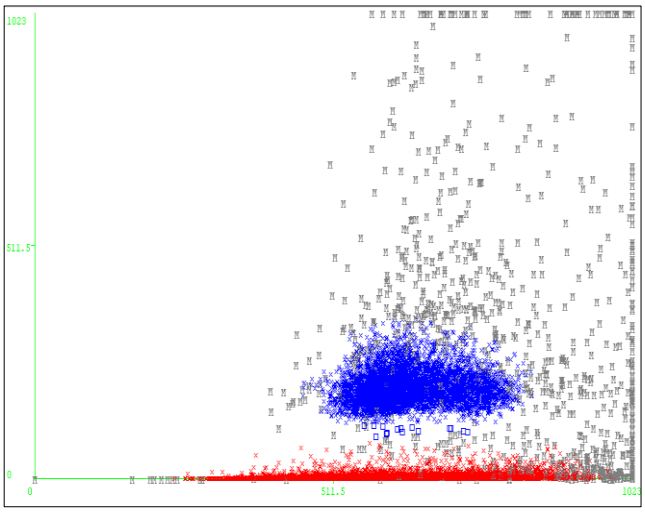


Figura 54: Risultato del clustering (SSL-PI)

3.3.3 Identificazione del miglior valore dei parametri

Basandosi sui risultati ottenuti con il data set Z0019566 si sono voluti individuare i migliori valori per i parametri caratteristici di DBScan.

Il valore di **ε** e il numero minimo di punti **minPoints** all'interno della sfera rappresentano i parametri da ottimizzare tenendo in considerazione non solo la corretta identificazione degli

spermatozoi, ma anche le cellule che vengono classificate come rumore e quindi non vengono inserite in nessun cluster.

Si è cercato di ottimizzare i parametri affinché il cluster degli spermatozoi fosse unico e con un valore di recall superiore al 90%.

minPoints	ε	# cluster	# oggetti noise	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	Cn
					S	U	S	U				
3	0,02	56	874	0,00472802	0,999	0,001	0,922	0,0001	S	5434	0	-
									U	2	14559	-
3	0,03	47	435	0,033812758	0,996	0,004	0,959	0,0013	S	0	5651	-
									U	14736	20	-
30	0,03	2	1423	0,010733427	0,999	0,001	0,909	0,0003	S	0	5356	
									U	14322	5	

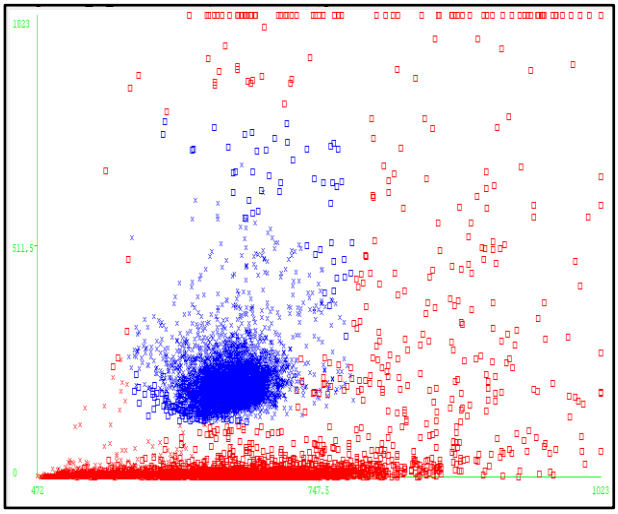


Figura 55: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

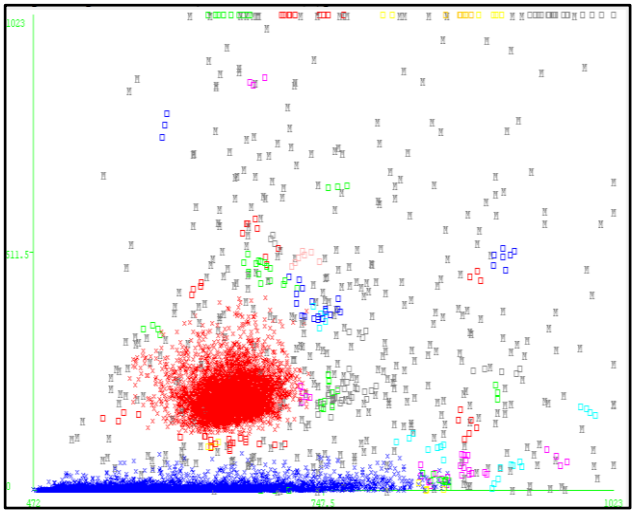


Figura 56: Risultato del clustering (FSL-PI).Epsilon = 0.03, minPoints = 3

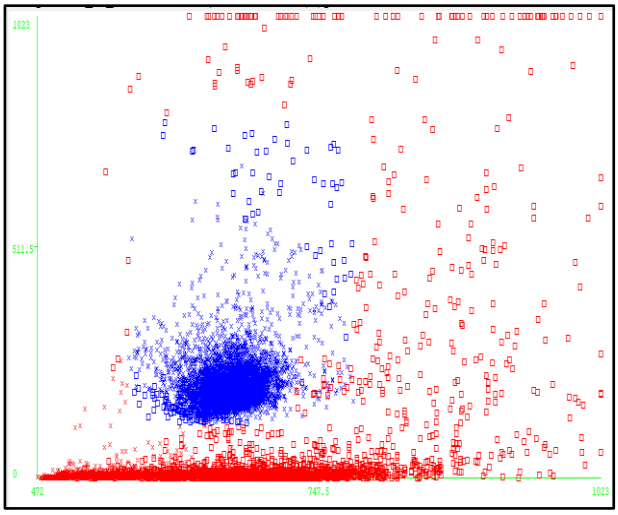


Figura 57: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

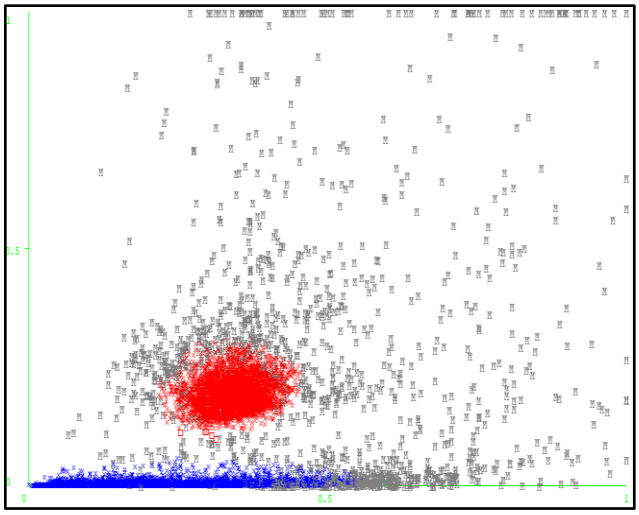


Figura 58: Risultato del clustering (FSL-PI).Epsilon = 0.03, minPoints = 30

Si sono ottenuti valori di recall superiori al 90% utilizzando **ε = 0.02** e valori di **minPoints** compresi tra [6..3] con il miglior valore di recall pari a 0.922 per **minPoints = 3**.

Valori di recall superiori al 90% si sono ottenuti anche con **ε = 0.03** e **minPoints** compreso tra [3..30]. Con **minPoints=30** si sono ottenuti due soli cluster di cui quello degli spermatozoi con un valore di recall pari a 0.909. Per **minPoints = 3** il valore di recall è stato pari a 0.959 indentificando però 47 cluster.

3.3.4 Validazione dei parametri ottimali con altri data set

Identificati i valori ottimali per i parametri di DBScan, l'algoritmo è stato testato su altri due campioni per verificare le sue capacità di generalizzazione.

Per poter eseguire una valutazione il più possibile attendibile sono stati scelti due data set acquisiti con le stesse impostazioni del citofluorimetro del campione utilizzato nella fase esplorativa. I campioni utilizzati sono: Z0019565 e Z0019561.

3.3.4.1 Campione Z0019565

minPoints	ε	# cluster	# oggetti noise	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	Cn
					S	U	S	U				
3	0,02	87	894	0,020111924	0,998	0,002	0,947	0,001	S	0	9867	-
									U	17887	19	-
30	0,03	2	1701	0,006350155	0,999	0,001	0,935	0,0002	S	0	9735	
									U	17603	5	

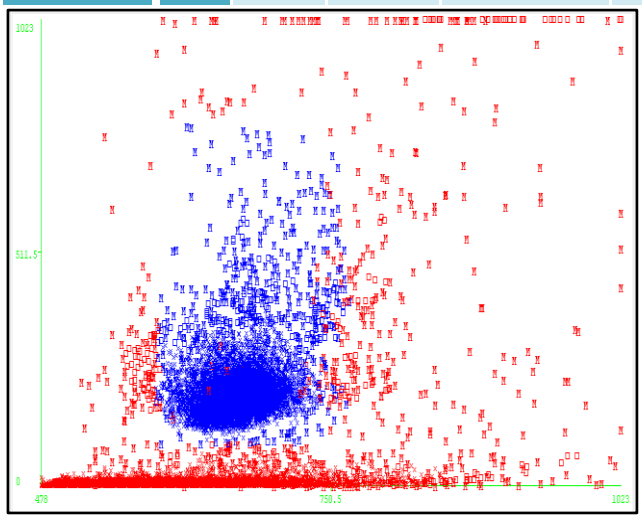


Figura 59: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

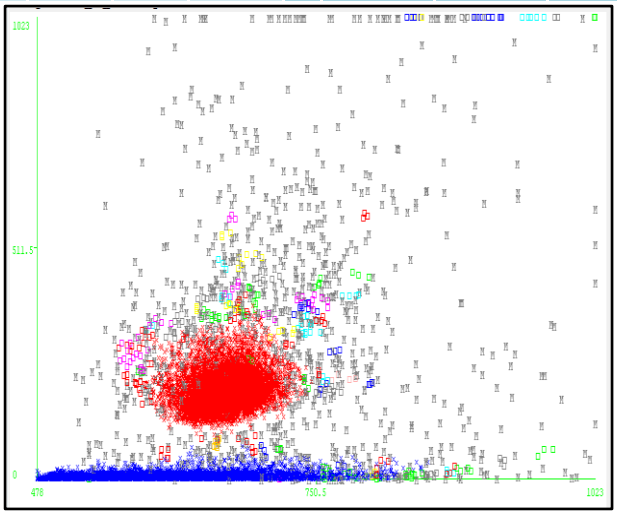


Figura 60: Risultato del clustering (FSL-PI).Epsilon = 0.02,minPoints = 3

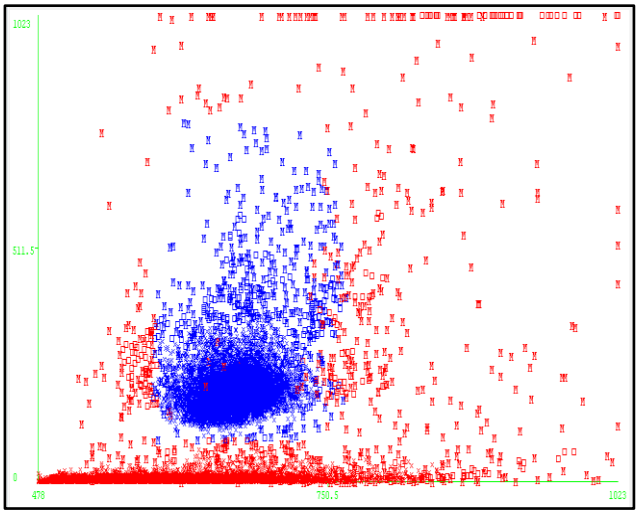


Figura 61: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

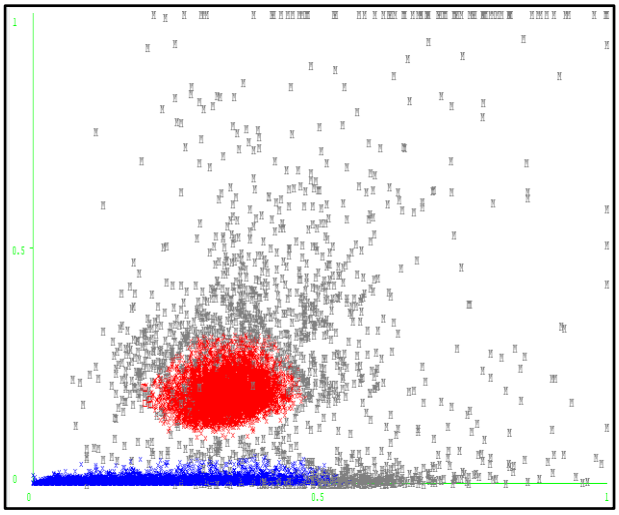


Figura 62: Risultato del clustering (FSL-PI).Epsilon = 0.03,minPoints = 30

3.3.4.2 Campione Z0019561

minPoints	ε	# cluster	# oggetti noise	Entropia Cluster Spermatozoi	Purezza Cluster Spermatozoi		Recall Cluster Spermatozoi			C0	C1	Cn
					S	U	S	U				
3	0,02	117	1161	0,019381798	0,998	0,002	0,929	0,0024	S	0	9758	-
									U	5906	18	-
30	0,03	2	2705	0	1	0	0,913	0	S	0	9585	
									U	5550	0	

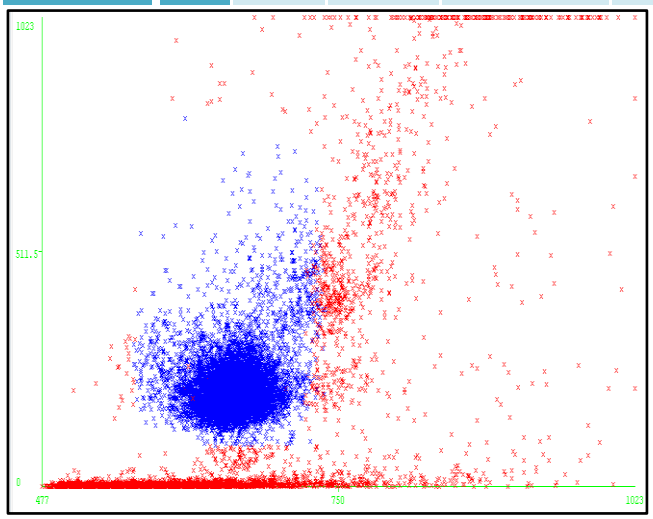


Figura 63: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

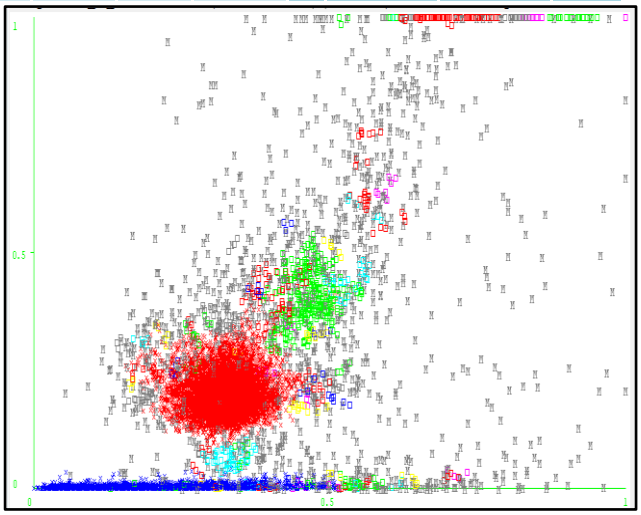


Figura 64: Risultato del clustering (FSL-PI).Epsilon = 0.02,minPoints = 3

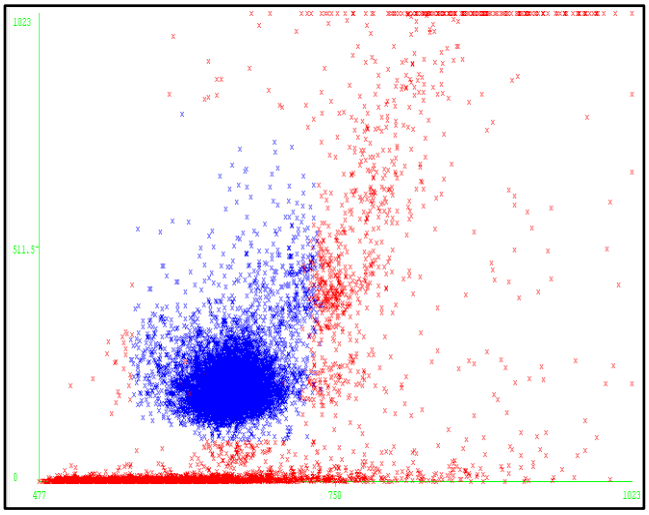


Figura 65: Classificazione dell'esperto del dominio (FSL-PI)). In blu gli spermatozoi, in rosso le altre cellule

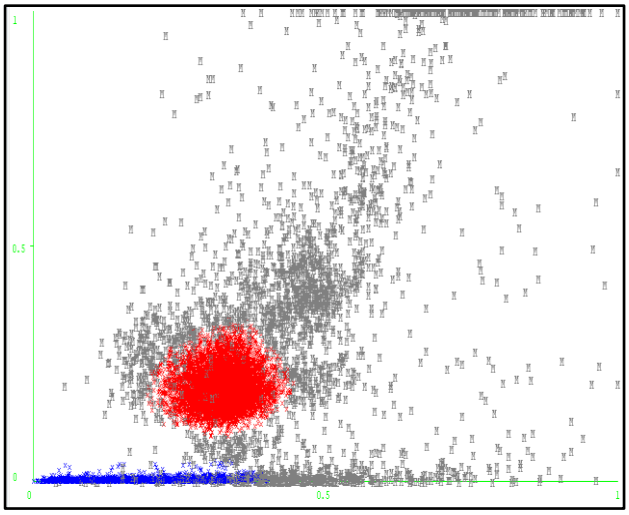


Figura 66: Risultato del clustering (FSL-PI).Epsilon = 0.03,minPoints = 30

3.3.4.3 Considerazioni

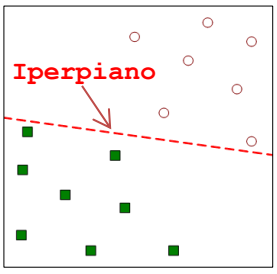
Su entrambi i campioni l'algoritmo DBScan conferma il buon risultato ottenuto nell'analisi esplorativa. I risultati migliori in termini di recall si sono ottenuti con **ε = 0.02** e **minPoints = 3** individuando però decine di cluster. Con **ε = 0.03** e **minPoints = 30** si sono ottenuti comunque ottimi risultati identificando però solo due cluster, uno dei quali molto rappresentativo degli spermatozoi (recall medio pari a 0.92%).

Cap. 4 - Classificazione

Alla luce dei risultati conseguiti con le tecniche di clustering si è voluta testare una tecnica di classificazione supervisionata per confrontarne i risultati.

La classificazione solitamente è un processo in due fasi nel quale l'obiettivo finale è l'individuazione di un attributo discreto, la classe. Nella prima fase si costruisce un modello che lega gli attributi delle istanze del data set al valore della relativa classe. Nella seconda fase si utilizza il modello appreso, sotto forma di regole di classificazione o formule matematiche, per classificare nuove istanze mai viste.

Le tecniche di classificazione sono diverse *[9]*: decision tree, rule based, neural network, support vector machines e naïve Bayes. Ogni tecnica impiega un algoritmo di apprendimento specifico per individuare un modello che meglio descrive la relazione tra gli attributi predittori e l'etichetta di classe delle istanze del data set.
Ogni tecnica ha un campo di applicazione specifico nel quale riesce ad ottenere generalmente risultati migliori delle altre. Le support vector machine, ad esempio, trattano naturalmente data set con attributi numerici. Poiché i dati provenienti dal citofluorimetro sono esclusivamente numerici e il problema è di classificazione binaria (spermatozoo/non-spermatozoo) si è deciso di testare un algoritmo di classificazione della famiglia delle support vector machines (SVM).
Le SVM cercano nello spazio dei parametri gli iperpiani a massimo margine che separano i dati di esempio delle diverse classi (problema di ottimizzazione vincolata). Le SVM sono utilizzabili anche nel caso in cui le classi non fossero linearmente separabili. L'accorgimento adottato in questo caso è di trasformare i dati di input con un mapping non lineare, in modo tale che nel nuovo spazio il confine tra le classi sia nuovamente rappresentabile con un iperpiano. Nella pratica, per questioni di efficienza computazionale, si ricorre a kernel function per il calcolo del modello *[9][11]*.



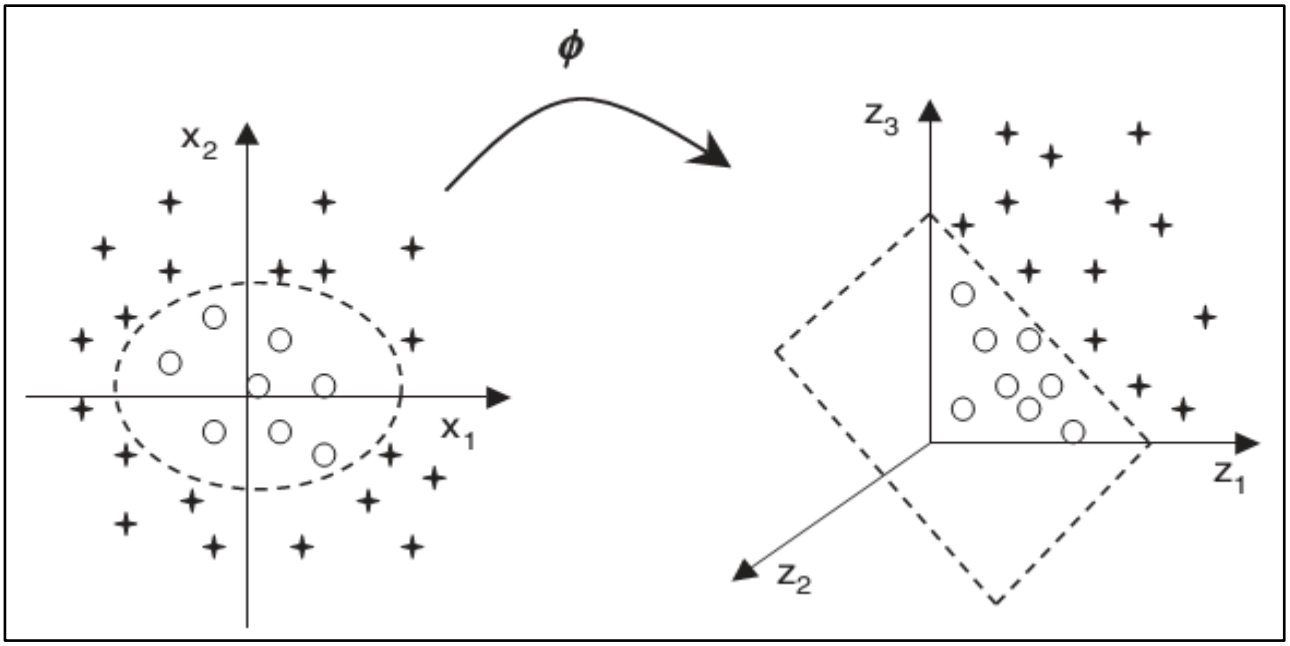


Figura 67: Mapping di uno spazio di R2 in uno spazio di R3. Il confine (decision boundary) a forma di ellisse nella immagine di sinistra diventa un iperpiano dopo avere eseguito il mapping in uno spazio dei parametri con una funzione non lineare. (figura presa da: B. Schölkopf, A.J. Smola, Learning with Kernels, MIT press, Cambridge,2002)

Nello specifico, per addestrare un classificatore basato su support vector è stata utilizzata una implementazione dell'algoritmo sequential minimal-optimization SMO (Platt, 1998; Keerthi et al., 2001) che fa uso di kernel polinomiali o Gaussiani per risolvere in maniera efficiente il problema di ottimizzazione.

Per poter confrontare correttamente i risultati è stato utilizzato lo stesso data set impiegato per gli algoritmi di clustering: campione Z0019566 etichettato manualmente dall'esperto.
Anche in questo caso dall'analisi sono state escluse le *feature* AUX e FITC come motivato nel Cap. 3 - . Inoltre l'algoritmo di classificazione è stato applicato esclusivamente alle *feature* FSLog, SSLog e PI poiché dall'analisi dei risultati del clustering si è dimostrato che queste tre *feature* sono quelle che hanno permesso una migliore identificazione degli spermatozoi.

4.1 Sequential Minimal Optimization

L'algortimo SMO è stato applicato al data set modificando la funzione di **kernel** e valutando i risultati con cross-validazione suddividendo il data set in 10 fold. Come valore per il parametro C, che rappresenta la complessità, è stato scelto quello che ha fornito le

migliori prestazioni per ogni kernel[19] *[13]*. Per tutti i kernel i risultati migliori si sono avuti con C = 1.

4.1.1 *Feature* FSLog - SSLog - PI

4.1.1.1 Kernel

kernel		Predetto S	Predetto U	TP rate	FP rate	Precisione	Recall	F-measure
PolyKernel[20]	S	5891	0	1	0,021	0,948	1	0,973
	U	323	14892	0,979	0	1	0,979	9,989
	Media pesata			0,985	0,006	0,985	0,985	0,989
	Istanze classificate correttamente					20783		98.4696 %
	Istanze classificate incorrettamente					323		1.5304 %
	Statistica Kappa					96,26 %		
Normalized PolyKernel[21]	S	5875	16	0,997	0,013	0,967	0,997	0,982
	U	199	15016	0,987	0,003	0,999	0,987	0,993
	Media pesata			0,99	0,006	0,99	0,99	0,99
	Istanze classificate correttamente					20891		98,9813 %
	Istanze classificate incorrettamente					215		1,0187 %
	Statistica Kappa					97,49 %		
Puk[22]	S	5881	10	0,998	0,003	0,991	0,998	0,995
	U	52	15163	0,997	0,002	0,999	0,997	0,998
	Media pesata			0,997	0,002	0,997	0,997	0,997
	Istanze classificate correttamente					21044		99,7062 %
	Istanze classificate incorrettamente					62		0,2938 %
	Statistica Kappa					99,27 %		

[19] È stata utilizzata la funzione di Weka "weka.classifiers.meta.CVParameterSelection -P "C 0.1 100.0 10.0" -X 5 -S 1 -W weka.classifiers.functions.SMO" variando il parametro C nel range [0,1..100]

[20] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

[21] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.NormalizedPolyKernel -C 250007 -E 2.0"

[22] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.Puk -C 250007 -O 1.0 -S 1.0"

kernel		Predetto S	Predetto U	TP rate	FP rate	Precisione	Recall	F-measure
RBF Kernel[23]	S	5891	0	1	0,02	0,95	1	0,974
	U	310	14905	0,98	0	1	0,98	0,99
	Media pesata			0,985	0,006	0,986	0,985	0,985
	Istanze classificate correttamente					20769		98,5312 %
	Istanze classificate incorrettamente					310		1,4688 %
	Statistica Kappa					96,41 %		

4.1.2 Considerazioni sulla terna di *feature* FSLog-SSLog-PI

Tra i diversi kernel testati, il PUK (kernel universale basato sulla funzione di Pearson VII *[14]*) ha permesso di identificare il maggior numero di spermatozoi includendo un numero molto limitato di cellule unknown.

4.1.3 Validazione dei parametri ottimali con altri data set

Appreso il modello del classificatore, questo è stato valutato sui due campioni per verificare le sue capacità di generalizzazione.

Per poter eseguire una valutazione il più possibile attendibile sono stati scelti due data set acquisiti con le stesse impostazioni del citofluorimetro del campione utilizzato nella fase esplorativa. I campioni utilizzati sono: Z0019565 e Z0019561.

4.1.3.1 Campione Z0019565

kernel		Predetto S	Predetto U	TP rate	FP rate	Precisione	Recall	F-measure
PolyKernel[24]	S	10403	6	0,999	0,024	0,959	0,999	0,979
	U	443	18192	0,976	0,001	1	0,976	0,988
	Media pesata			0,985	0,009	0,985	0,985	0,985
	Istanze classificate correttamente					28595		98,4541 %
	Istanze classificate incorrettamente					449		1,5459 %
	Statistica Kappa					96,67 %		

[23] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"

[24] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

kernel		Predetto S	Predetto U	TP rate	FP rate	Precisione	Recall	F-measure
Normalized PolyKernel[25]	S	10395	14	0,999	0,017	0,97	0,999	0,984
	U	326	18309	0,983	0,001	0,999	0,983	0,991
	Media pesata			0,988	0,007	0,989	0,988	0,991
	Istanze classificate correttamente					28704		98,8294 %
	Istanze classificate incorrettamente					340		1,1706 %
	Statistica Kappa					97,47 %		
Puk[26]	S	10403	6	0,999	0,01	0,982	0,999	0,99
	U	194	18441	0,99	0,001	1	0,99	0,995
	Media pesata			0,993	0,004	0,993	0,993	0,993
	Istanze classificate correttamente					28844		99,3114 %
	Istanze classificate incorrettamente					200		0,6886 %
	Statistica Kappa					98,51 %		
RBF Kernel[27]	S	10391	18	0,998	0,024	0,959	0,998	0,978
	U	442	18193	0,976	0,002	0,999	0,976	0,988
	Media pesata			0,984	0,01	0,985	0,984	0,984
	Istanze classificate correttamente					28584		98,4162 %
	Istanze classificate incorrettamente					460		1,5838 %
	Statistica Kappa					96,59 %		

[25] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.NormalizedPolyKernel -C 250007 -E 2.0"

[26] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.Puk -C 250007 -O 1.0 -S 1.0"

[27] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"

4.1.3.2 Campione Z0019561

kernel		Predetto S	Predetto U	TP rate	FP rate	Precisione	Recall	F-measure
PolyKernel[28]	S	10476	17	0,998	0,134	0,914	0,998	0,932
	U	988	6359	0,866	0,002	0,997	0,866	0,932
	Media pesata			0,944	0,08	0,948	0,944	0,943
	Istanze classificate correttamente					16835		94,3666 %
	Istanze classificate incorrettamente					1005		5,6334 %
	Statistica Kappa					88,14 %		
Normalized PolyKernel[29]	S	10479	14	0,999	0,095	0,937	0,999	0,952
	U	701	6646	0,905	0,001	0,998	0,905	0,949
	Media pesata			0,96	0,057	0,962	0,96	0,96
	Istanze classificate correttamente					17125		95,9922 %
	Istanze classificate incorrettamente					715		4,0078 %
	Statistica Kappa					91,61 %		
Puk[30]	S	10484	9	0,999	0,07	0,953	0,999	0,976
	U	513	6834	0,93	0,001	0,999	0,93	0,963
	Media pesata			0,971	0,041	0,972	0,971	0,971
	Istanze classificate correttamente					17318		97,074 %
	Istanze classificate incorrettamente					522		2,926 %
	Statistica Kappa					93,9 %		
RBF Kernel[31]	S	10454	39	0,996	0,135	0,913	0,996	0,953
	U	990	6357	0,865	0,004	0,994	0,865	0,925
	Media pesata			0,942	0,081	0,947	0,942	0,931
	Istanze classificate correttamente					16811		94,2321 %
	Istanze classificate incorrettamente					1029		5,7679 %
	Statistica Kappa					87,86		

[28] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

[29] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.NormalizedPolyKernel -C 250007 -E 2.0"

[30] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.Puk -C 250007 -O 1.0 -S 1.0"

[31] weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"

4.1.3.3 Considerazioni

Il kernel PUK anche nei campioni di test ha avuto i migliori risultati garantendo la corretta individuazione di gran parte degli spermatozoi a fronte però di un numero di falsi positivi relativamente elevato.

Cap. 5 - Software

Questo lavoro ha richiesto l'utilizzo di alcuni software adottati in ambito biomedicale e lo sviluppo di alcune applicazioni specifiche per gli obiettivi prefissati.

La prima fase del progetto ha riguardato lo studio e la comprensione delle proprietà delle cellule di un eiaculato. Per fare ciò ci si è avvalsi di un tool, Weasel, che solitamente gli esperti del dominio utilizzano per analizzare i dati del campione di eiaculato ottenuti dal suo passaggio nel citofluorimetro.

Nella seconda fase del progetto si sono applicate tecniche di data mining alle informazioni ricavate con il citofluorimetro. Prerequisito a questa seconda fase è stato riuscire a collezionare le informazioni del campione in un formato idoneo per essere in seguito elaborate. Il primo tool realizzato è stato quindi allo scopo di interpretare il file del citofluorimetro e memorizzarne le informazioni in un formato interpretabile dallo strumento di data mining. Durante la fase esplorativa ci si è avvalsi di un software di data mining specifico utilizzando la sua interfaccia grafica. Una volta individuato l'algoritmo più adatto si è sviluppato un software, basato su quella tecnica, che attraverso un workflow dall'output del citofluorimetro all'identificazione degli spermatozoi potesse aiutare l'esperto del dominio nelle analisi dell'eiaculato.

5.1 Weasel

Weasel è un programma sviluppato presso la WEHI (Walter+Eliza Hall – Institute of Mediacal Research) per l'analisi e la visualizzazione di dati provenienti dalla citofluorimetria a flusso. Come è possibile vedere nell'immagine sottostante, sono disponibili differenti formati di visualizzazione dai quali si possono estrarre informazioni numeriche e statistiche anche provenienti da campioni diversi contemporaneamente. Tra le varie funzioni è anche possibile selezionare una parte degli oggetti visualizzati, eseguendo un gate sul campione, e salvare il valore delle *feature* delle sole cellule del gate in un file compatibile con lo standard FCS.

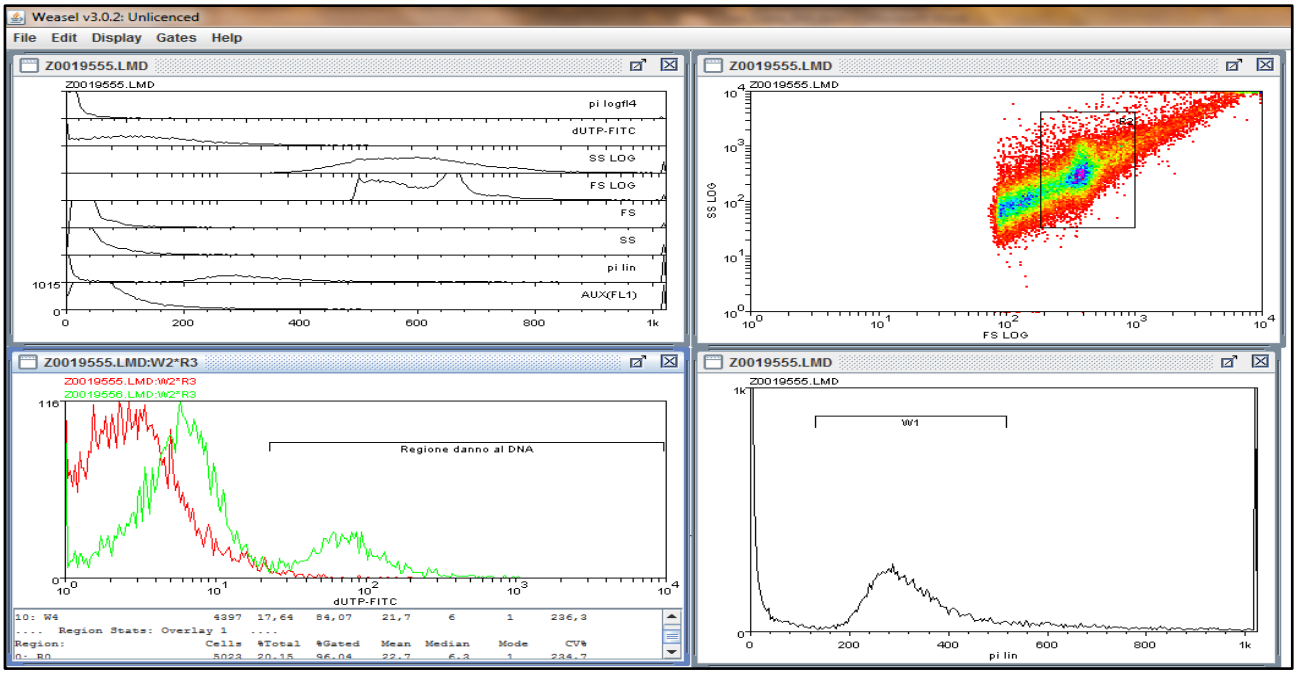


Figura 68: Tool Weasel

Questo strumento è utilizzato dagli esperti del dominio per leggere il file output del citofluorimetro e visualizzare i dati collezionati sulle cellule. Attraverso un processo di ispezione del campione vengono individuati gli spermatozoi come illustrato nel paragrafo "1.5 Citofluorimetria - Analisi manuale dello spermiogramma".

5.2 FSC Process – Analisi Citofluorimetrica

Dall'esperienza acquisita durante la fase esplorativa si è ritenuta più attuabile l'adozione di una tecnica di data mining non supervisionata in quanto la disponibilità di dati correttamente etichettati, essendo un processo particolarmente lungo e laborioso, non è garantita dall'esperto del dominio. Nello specifico si è deciso di implementare una soluzione basata su DBScan poiché confrontando i risultati ottenuti nella fase esplorativa, si è ritenuta la più promettente.

Il processo di sviluppo del software ha richiesto quindi l'implementazione di alcune applicazioni Java-based che applicate in cascata avessero come effetto ultimo l'identificazione di uno schema di clustering nel quale ogni cluster fosse etichettato come "insieme di spermatozoi" oppure come "insieme di cellule di altro genere".

Nella Figura 69 è rappresentata la suddivisione delle varie classi del progetto tra i diversi package:

- “data” e “fcs2converter” contengono le classi utilizzate nella prima fase del progetto. Queste permettono di interpretare il file output del citofluorimetro e il file generato con Weasel dopo che l’esperto del dominio ha etichettato le diverse cellule con un gate sugli spermatozoi.
- “mining.classificazione” e “mining.clustering” contengono le implementazioni delle funzioni di data mining
- “util” contiene classi di supporto
- “main” contiene i main utilizzati per le diverse parti del progetto

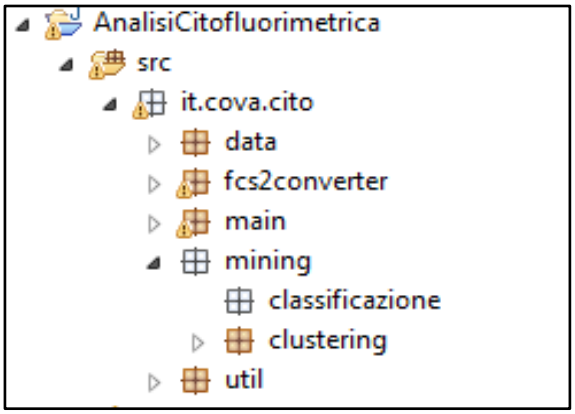


Figura 69: Struttura del progetto "Analisi Citofluorimetrica"

5.2.1 FCS Converter

Il primo passo per arrivare ad analizzare le informazioni sulle cellule del campione con tecniche di data mining ha visto la realizzazione di un software che fosse in grado di leggere il file .LMD output del citofluorimetro. Questo file raccoglie tutte le informazioni sul campione analizzato e le memorizza secondo lo standard FCS2.0, v. paragrafo 2.1.

La classe *Converter* ha il compito di leggere il file del citofluorimetro e interpretarlo correttamente. A questo scopo individua le sezioni HEADER, TEXT, DATA e ANALYSIS. Ogni campione è rappresentato da un *Experiment* composto da un insieme di *Cell* che rappresentano le misurazioni di ogni cellula del campione stesso.

Nella sezione TEXT si individuano le diverse keyword con i relativi valori e nella sezione DATA si estraggono i valori delle *feature* misurate per ogni cellula.

L’oggetto *Experiment*, dopo aver analizzato il file .LMD, permette di salvarne le informazioni in diversi formati testuali, uno dei quali aderente al formato ARFF (Attribute-

Relation File Format). È stato scelto il formato ARFF poiché nella fase successiva del progetto sono state utilizzate le librerie del software di data mining Weka che nativamente supporta questa rappresentazione dei dati.

5.2.2 Addestramento, etichettatura e validazione

L'esperto del dominio per etichettare le cellule del campione come spermatozoo o come cellule di altro genere utilizza il tool Weasel. Identificati gli spermatozoi, esporta questo insieme di cellule in un nuovo data set, un file di dati FCS. La classe *WeaselConverter* ha il compito di trasformare le informazioni contenute in questo file nel formato ARFF.
Successivamente per ottenere l'intero data set con tutte le cellule etichettate, la classe *Etichettatore* ha il compito di confrontare i due file e aggiungere un ulteriore attributo, CLASS, che indica la classe di appartenenza di tutte le cellule del campione tra SPERMATOZOO e UNKNOWN.

Una volta trasformati i file del citofluorimetro ed etichettate le cellule del campione, sono state applicate le tecniche di data mining come descritto in "Cap. 3 - Clustering" e "Cap. 4 - Classificazione". Nella fase esplorativa sono state valutate diverse tecniche, utilizzando il tool di data mining Weka [32]*[16]*.
Per la valutazione degli schemi di clustering si è utilizzato un foglio di calcolo appositamente realizzato per il calcolo delle misure di entropia, purezza, precisione, recall e F-measure di ogni cluster individuato.

5.2.3 Clusterer

Conclusa la parte di addestramento di un classificatore e la valutazione delle tecniche di clustering si è individuata la tecnica di data mining ritenuta più idonea a questo progetto: DBScan. Si è implementata quindi una applicazione Java che appoggiandosi alle API di Weka potesse eseguire tale tecnica su data set nuovi per individuare gli spermatozoi all'interno del campione.
L'intero processo segue queste fasi: i file .LMD ottenuti analizzando nuovi campioni al citofluorimetro sono interpretati e trasformati in formato ARFF il quale è elaborato dalla

[32] Weka implementa una collezione di algoritmi di machine learning per compiti di data mining come pre-processing, classificazione, regressione, clustering e regole associative. Consente di utilizzare gli algoritmi direttamente dalla sua interfaccia grafica oppure all'interno di codice Java attraverso API specifiche.

classe *ClusteringWekaDBScan*. Questo oggetto si occupa di aprire il data set e avviare il clusterer DBScan con i parametri ottimali individuati durante l'analisi esplorativa. Segue poi la fase di riconoscimento del/dei cluster di spermatozoi secondo alcuni principi individuati nell'analisi esplorativa. Sono quindi etichettate tutte le cellule del campione con la classe SPERMATOZOO o UNKNOWN. Il risultato dell'individuazione degli spermatozoi è visualizzabile sia a video mediante una finestra grafica nella quale le cellule delle due classi sono rappresentate con colori differenti, oppure è possibile memorizzare il risultato dell'etichettatura in diversi file:

- data set con tutte le cellule etichettate SPERMATOZOO/UNKNOWN
- data set con tutte le cellule etichettate in base al cluster di appartenenza
- data set con solo le cellule identificate come spermatozoi

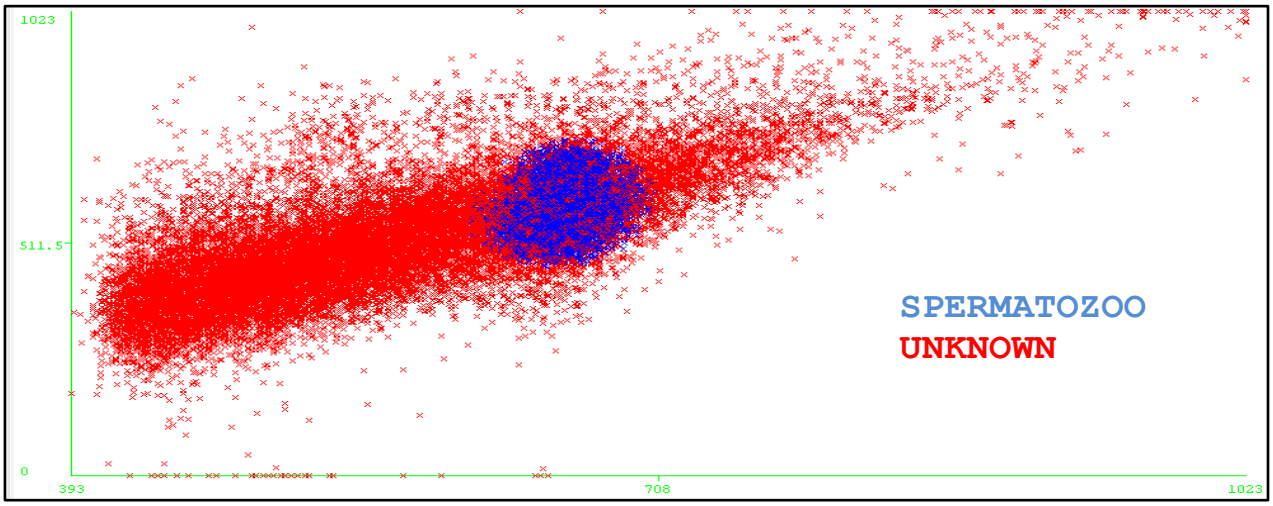


Figura 70: risultato grafico del clustering

```
@relation Dati_Citofluorimetrici__SE_7090_T__Z0009264-Classificazione

@attribute PIL numeric
@attribute FITC numeric
@attribute SSL numeric
@attribute FSL numeric
@attribute FS numeric
@attribute SS numeric
@attribute PI numeric
@attribute AUX numeric
@attribute CLASS_OLD {SPERMATOZOO,UNKNOWN}
@attribute CLASS {SPERMATOZOO,UNKNOWN}

@data

0,0,336,447,5,2,0,9,?,UNKNOWN
0,0,278,421,4,1,0,2,?,UNKNOWN
318,0,550,664,40,14,96,26,?,SPERMATOZOO
0,0,344,484,7,2,0,4,?,UNKNOWN
0,0,322,427,4,1,3,15,?,UNKNOWN
278,393,598,626,28,22,55,661,?,UNKNOWN
97,183,591,559,15,20,12,127,?,UNKNOWN
419,21,589,668,41,20,213,78,?,SPERMATOZOO
0,147,827,800,136,174,9,111,?,UNKNOWN
411,170,510,649,35,10,189,119,?,SPERMATOZOO
438,207,567,671,43,16,228,154,?,SPERMATOZOO
0,40,461,571,17,6,11,57,?,UNKNOWN
336,455,727,675,44,71,93,1023,?,UNKNOWN
0,0,419,525,11,4,10,22,?,UNKNOWN
```

Figura 71: risultato testuale del clustering

Cap. 6 - Conclusioni e sviluppi futuri

L'analisi di dati provenienti dalla citofluorimetria a flusso è un processo estremamente time-consuming. Nell'analisi di campioni di eiaculato la procedura manuale prevede di visualizzare i valori di alcune caratteristiche fisiche delle cellule e individuare gli spermatozoi secondo una procedura non standard e difficilmente riproducibile.

In prima battuta lo scopo di questo lavoro si è orientato alla verifica e validazione delle *feature* utilizzate dagli esperti del dominio in questo tipo di esami. Non esiste infatti un accordo tra i biologi che sancisce quali *feature* bisogna utilizzare per individuare al meglio gli spermatozoi, ogni team utilizza l'insieme che ritiene più opportuno. I test hanno dimostrato che i migliori risultati in termini di individuazione di spermatozoi, mantenendo limitato il numero di falsi positivi, si sono avuti nell'utilizzo delle *feature* misurate in scala logaritmica per quanto riguarda i parametri fisici delle cellule, mentre per quanto riguarda la misura della fluorescenza del DNA i risultati migliori si sono registrati con il valore di Propidio Ioduro in scala lineare. Inoltre, a conferma della correttezza della metodologia adottata dal Prof. Davide Bizzaro e del suo team di biologi nell'analisi manuale dello spermiogramma, vedi paragrafo 1.5, l'utilizzo combinato delle *feature* FSL, SSL e PI ha fornito i risultati più soddisfacenti.

Individuato l'insieme di *feature* migliori, il lavoro è proseguito nella direzione di sperimentare diverse tecniche di data mining alla ricerca di quale fosse più efficace nell'isolare gli spermatozoi. Nel valutare le diverse tecniche si è tenuto conto non solo della sensitività, cioè di quanti spermatozoi si individuano sul numero totale presente nel campione, ma anche della specificità che dà una indicazione del numero di falsi negativi rilevati.

Le tecniche testate sono state sia supervisionate che non supervisionate. Poiché però in generale è sempre difficoltoso ottenere dati di buona qualità in quanto l'etichettatura manuale da parte dell'esperto è un processo lento e costoso, si sono testate con maggiore attenzione le tecniche non supervisionate. In particolare ne sono state scelte tre basate su principi completamente differenti l'uno dall'altro (metodi partitivi, basati sulla densità e

basati su modelli parametrici) per ottenere un'analisi più approfondita del problema e delle possibili soluzioni.

Tra le tecniche non supervisionate adottate, il clustering con K-Means è quella che ha ottenuto i risultati peggiori. Per quanto abbia individuato un buon numero di spermatozoi durante l'analisi esplorativa mantenendo anche un buon valore di specificità, non è riuscito a generalizzare efficacemente applicato ad altri data set: i cluster individuati sono sempre stati caratterizzati da bassi valori di sensitività. Nel caso in cui gli spermatozoi si dividessero in più cluster, il valore complessivo di sensitività sarebbe allora elevato ma bisognerebbe pagare un prezzo molto elevato in termini di falsi positivi.

L'algoritmo Expectation Maximisation ha ottenuto risultati migliori, seppure non ottimi. È sempre stato in grado di individuare gran parte degli spermatozoi all'interno del campione, includendo però anche un discreto numero di falsi positivi. Probabilmente la motivazione è da ricercare nel principio a cui si ispira: è una tecnica basata su un modello parametrico nel quale si ipotizza che i dati seguano un certo modello statistico, in questo caso un insieme di distribuzioni di probabilità Gaussiane. I risultati ottenuti evidenziano quindi che il modello supposto non rispecchia appieno la realtà dei dati, probabilmente a causa della loro variabilità molto ampia.

Una tecnica di clustering non basata su modelli statistici né influenzata dalla variabilità dei dati è DBScan. Al contrario delle tecniche sperimentate fino ad ora, DBScan utilizza un concetto di densità per individuare i cluster. Questo ha il vantaggio di individuare cluster di forma arbitraria e in numero non predeterminato. I risultati ottenuti sia con i dati di test che con nuovi data set sono stati inferiori, rispetto alle altre tecniche sperimentate, in termini di sensitività attestandosi tra il 90% e il 94%.

Questi risultati richiedono però una considerazione più approfondita. I valori di sensitività indicati precedentemente si sono ottenuti costantemente sia durante l'analisi esplorativa che nel momento in cui si è valutata la capacità dell'algoritmo di generalizzare. Le altre tecniche di clustering, al contrario, hanno ottenuto risultati migliori applicati ai dati di test, ma la loro capacità di generalizzare non è stata altrettanto buona perdendo in sensitività circa 4 punti percentuali nel caso di EM (passando da 99% a 95%) ed addirittura tra 4 e 27 punti nel caso di K-Means (passando da 99% a 95% e da 99% a 72% nel caso peggiore).

A favore dell'algoritmo DBScan vi è anche un'altra considerazione: nell'analisi dello spermiogramma il fine ultimo del medico è valutare il danno al DNA degli spermatozoi per

indirizzare il paziente verso la cura più idonea. Identificare un'alta percentuale di spermatozoi è sicuramente una necessità, a patto però che non siano incluse troppo cellule di altro genere. In questo caso il rischio è di quantificare il danno agli spermatozoi, che si misura in termini di quanti sono danneggiati e quanto lo sono, in maniera errata perché influenzato da cellule che non sono spermatozoi.
DBScan ha ottenuto risultati che sono andati proprio in questa direzione: la percentuale di spermatozoi individuati è stata sempre maggiore del 90% garantendo al contempo un numero di falsi positivi estremamente ridotto.
Un problema però potrebbe verificarsi con questo algoritmo: l'individuazione del parametro ottimale **MinPoints**. Nella fase esplorativa si è individuato il valore ottimale su un data set con 21106 istanze e lo si è validato, con successo, su data set con 29044 (+37.6%) e 17840 (-16,5%) istanze. Nel caso di data set con un numero di istanze molto diverso, il valore di **MinPoints** individuato potrebbe non essere altrettanto efficace. Bisogna quindi prendere in considerazione la possibilità di modificare proporzionalmente il suo valore sulla base del numero totale di oggetti presenti nel campione.

Conclusa l'analisi delle tecniche di clustering si è testata la bontà di una tecnica di classificazione supervisionata basata sulle support vector machines: SMO. I risultati ottenuti sono stati molto interessanti: il modello individuato riapplicato al training set ha garantito una percentuale di istanze correttamente etichettate pari al 99,8% (valutazione con cross-validazione) con un tasso di falsi positivi per quanto riguarda gli spermatozoi di 0,3%.
Il modello appreso applicato a nuovi data set, com'era prevedibile ha mostrato un tasso totale di classificazione corrette in discesa rispetto al training set. Gli spermatozoi sono stati individuati praticamente sempre tutti (recall 99,9%), però si è riscontrato un notevole innalzamento della percentuale di cellule erroneamente classificate come spermatozoi, variando tra 1% e 7%.

Nel complesso è emerso che SMO è riuscito ad individuare praticamente tutti gli spermatozoi nei diversi data set pagando però il prezzo di un elevato numero di falsi positivi. DBScan invece è andato nella direzione opposta: non è stato in grado di individuare un numero di spermatozoi elevato come SMO, ma a fronte di un livello di sensitività mai inferiore al 90% ha garantito un FP-rate molto basso attorno a 0,1%.

L’adozione di una tecnica di clustering pone però la questione del riconoscimento del/dei cluster di spermatozoi. A differenza di quanto accade con un algoritmo di classificazione, che etichetta le istanze con la classe di appartenenza, nel clustering le cellule sono suddivise in insieme coerenti, quindi il passo successivo è individuare quale tra questi insiemi contengono spermatozoi. Avendo a disposizione pochi campioni etichettati dall’esperto, le considerazioni qui esposte non possono considerarsi esaustive.
Dai campioni etichettati dall’esperto è emerso che l’insieme di spermatozoi tende ad assumere una forma sferica o ellissoidale, e il suo valor medio e varianza, rispetto alle singole *feature*, segue un determinato andamento.
L’individuazione degli spermatozoi nei nuovi data set avviene quindi incrociando i dati sui cluster individuati, con le informazioni sopra descritte e facendo alcune considerazioni sui valori ammissibili della *feature* PI.

6.1 Ultime considerazioni

Da questo lavoro sono emersi due validi strumenti che possono aiutare il medio-biologo nell’analisi dello spermiogramma. DBScan predilige la qualità dell’insieme di spermatozoi, mentre SMO ottiene migliori risultati in termini di numero di individuazioni. La scelta definitiva dello strumento rimane ancora aperta perché deve essere fatta sul campo con l’aiuto dell’esperto del dominio.
Sono comunque stati individuati due strumenti che possono sostituire la classificazione manuale del medico-biologo con una automatica direttamente dai dati del citofluorimetro.

Il beneficio che potrebbe portare l’adozione di tecniche informatiche in questo tipo di studi si misurerebbe nella diminuzione dei tempi e costi di analisi, e nella riproducibilità dei risultati in quanto frutto di considerazioni matematiche e non legate all’intuizione, l’esperienza o lo stato d’animo dell’esperto del dominio. La conoscenza approfondita del dominio da parte dell’esperto è però stata fondamentale nella costruzione e nella validazione di modelli attendibili. Inoltre non bisogna dimenticare che l’analisi automatica dello spermiogramma non deve essere uno strumento che mira a sostituire le capacità

dell’esperto, ma deve andare nella direzione di coadiuvarlo semplificandone il lavoro e rendendolo meno soggetto ad errori.

Glossario

Attributo	vedi "*Feature*"
Campione	insieme di cellule ed altri organuli presenti nell'eiaculato di un pazione
Campione di controllo	parte del campione in cui è stata fatta avvenire la reazione biochimica attraverso la quale la molecola fluorescente FITC si lega alle strutture danneggiate del DNA.
Campione di test	parte del campione in cui non è stata fatta avvenire la reazione biochimica che permette di identificare il danno al DNA attraverso la fluorescenza della molecola FITC
Citofluorimetria a flusso	tecnica utilizzata per misurare e caratterizzare una popolazione di cellule sospese in un fluido secondo una serie di *feature*.
Citofluorimetro	strumento utilizzato per analizzare una popolazione di cellule con la tecnica della citofluorimetria a flusso
Data set	insieme di cellule di un campione. Ogni cellula è caratterizzata dal valore delle *feature* misurate.
Feature	proprietà fisica dell'oggetto analizzato
FITC	molecola fluorescente di Isotiocianato di Fluoresceina utilizzata per marcare le strutture di DNA danneggiate.
FS	forward scatter. *Feature* della cellula che contraddistingue la sua forma e dimensione
Gate	selezione elettronica di una parte di cellule di un campione sulla base di determinati parametri.
Marcatura aspecifica	molecola fluorescente che si lega alle cellule senza che sia presente del danno al DNA. Vedi "Campione di controllo"
Marcatura specifica	reazione biochimica attraverso la quale la molecola fluorescente di lega alle strutture danneggiate del DNA. Vedi "Campione di controllo"
PI	molecola fluorescente di Propidio Ioduro utilizzata per marcare il DNA presente nelle cellule.
SS	side scatter. *Feature* della cellula che contraddistingue la sua struttura interna.

Bibliografia

[1]	Ali Bashashanti, Ryan R. Brinkman, "A Survey of Flow Cytometry Data Analisys Methods", Hindawi Publishing Corporation, Advances in Bioinformatics, vol. 2009, Article ID 584603, 19 pages
[2]	http://cyto.purdue.edu/
[3]	http://www.ciam.unibo.it/
[4]	W. Roy Overton "Modified Histogram Subtraction Technique for Aanalysis of Flow Cytometry Data", Cytometry 9:619-626 (1988)
[5]	"Data File Standard for Flow Cytometry", Cytometry 11:323-332, 1990
[6]	Cytometry 5:553-555, 1984
[7]	"Making Sense of Data II – A Practical Guide to Data Visualization, Advanced Data Mining Methods and Application", GLENN J. MYATT WAYNE P. JOHNSON. Wiley
[8]	"Discovering Knowledge In Data - An Introduction to Data Mining", DANIEL T. LAROSE. Wiley
[9]	"Introduction to Data Mining", PANG-NING TAN, MICHAEL STEINBACH, VIPIN KUMAR. Pearson
[10]	"Introduction to Data Mining and its Applications", S. Sumathi, S.N. Sivanandam. Springer
[11]	"Data Mining Practical Machine Learning Tools and Techniques - Third Edition", Ian H. Witten, Eibe Frank, Mark A. Hall. Elsevier
[12]	"WEKA Manual", Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse. University of Waikato
[13]	http://weka.wikispaces.com/Optimizing+parameters
[14]	B. Uestuen, W.J. Melssen, L.M.C. Buydens (2006). Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. Chemometrics and Intelligent Laboratory Systems. 81:29-40
[15]	Y. D. Mahnke and M. Roederer, "Optimizing a multicolor immunophenotyping assay," Clinics in Laboratory Medicine,

	vol. 27, no. 3, pp. 469–485, 2007.
[15]	http://www.wehi.edu.au/faculty/advanced_research_technologies/flow_cytometry/weasel_for_flow_cytometry_data_analysis
[16]	http://www.cs.waikato.ac.nz/ml/weka/index.html