UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

Scuola Politecnica e delle Scienze di base

Area didattica Scienze Matematiche Fisiche e Naturali

Corso di Laurea in Informatica

Tecniche di Clustering basate sul Machine Learning

Tesi sperimentale di Laurea Triennale

Tutor Accademico
Prof. Roberto Prevete

Tutor Aziendale
Dr. Massimo Brescia

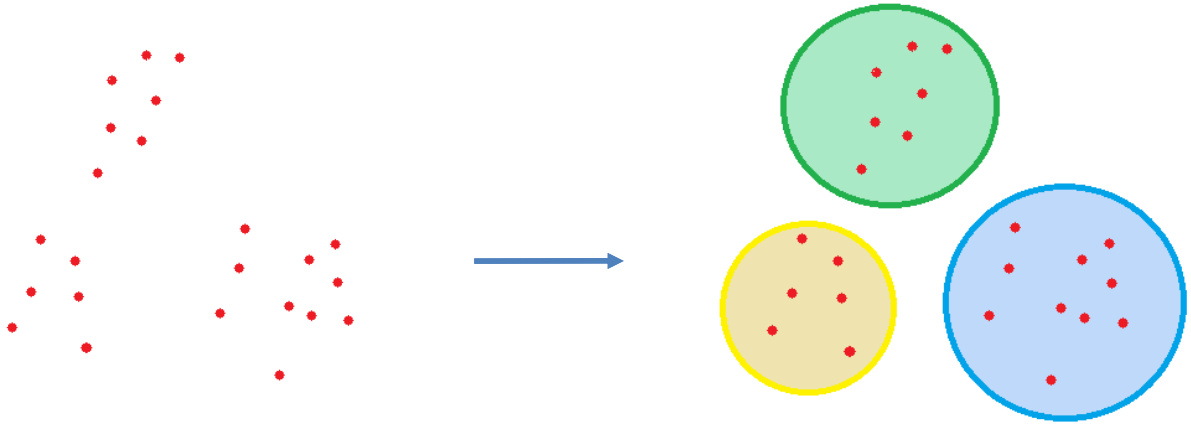
Candidato
Francesco Esposito
Matr. N86/24

Introduzione

Data **M**ining and **E**xploration **W**eb **A**pplication **RE**sources

Applicazione web dedicata al data mining tramite tecniche di machine learning

Estensione della suite DAMEWARE con modelli per clustering e riduzione dimensionalità

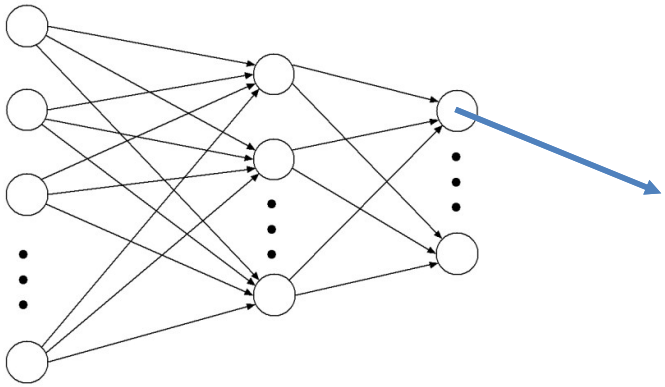


Clustering: raggruppamento di oggetti in base a criteri di similarità

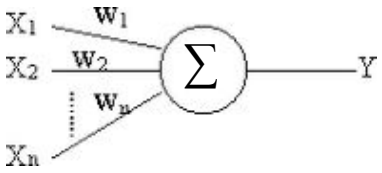
Indice dei contenuti

1. Introduzione alle reti neurali: modello SOM
2. Evoluzioni del modello SOM
   - Evolving SOM ed Evolving Tree
   - Two-Stage Clustering
3. Tecnologie utilizzate
4. Introduzione ai test: indici di qualità
   - Test 1: Iris (confronto con algoritmo di clustering standard: K-Means)
   - Test 2: Chainlink
   - Test 3: Target
   - Test 4: M101 (immagine astronomica a banda singola)
5. Conclusioni

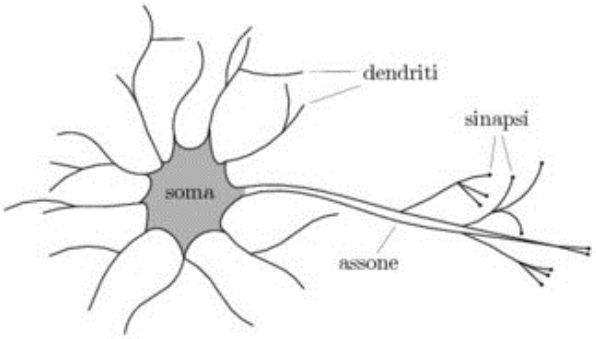
Introduzione alle reti neurali



Rete neurale



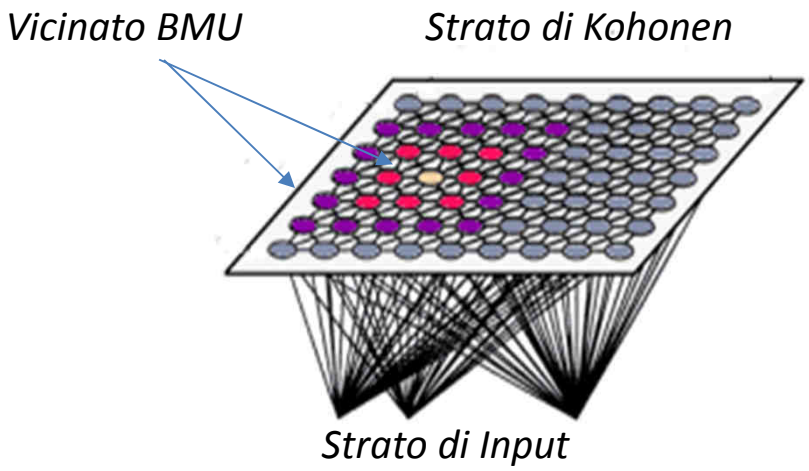
Neurone artificiale

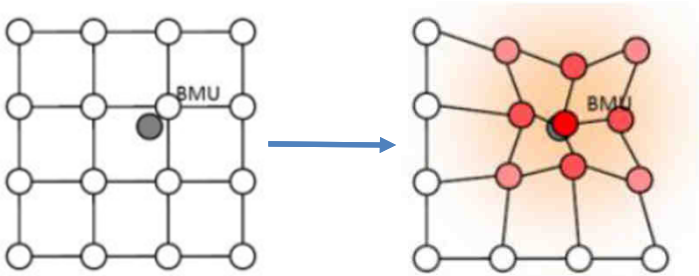


Neurone biologico

Il modello Self Organizing Map

Kohonen, T. (1990). The self-organizing map. Proceedings of IEEE, 78, 14641480





*Apprendimento secondo la regola **Winner Takes Most**: la gradazione di rosso indica il grado di apprendimento che diminuisce quanto più ci si allontana dal BMU*

Evoluzione topologia

Evoluzione apprendimento

SOM

Self Organizing Maps

TWL
Two Winners Linkage

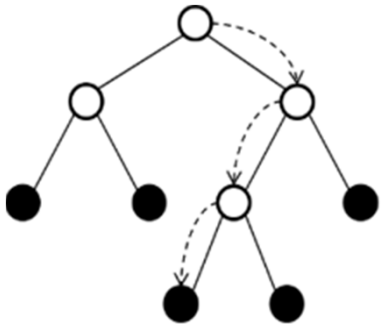
K-Means

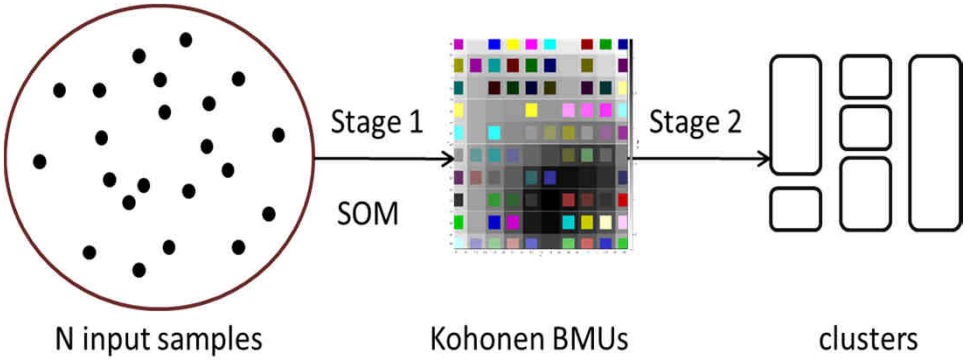
Umat-CC
U-matrix a Componenti Connesse

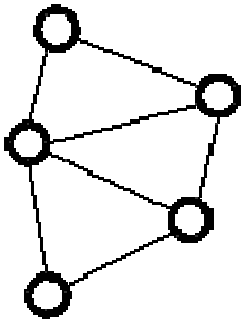
E-TREE
Evolving Tree

2-stage Clustering

E-SOM
Evolving SOM

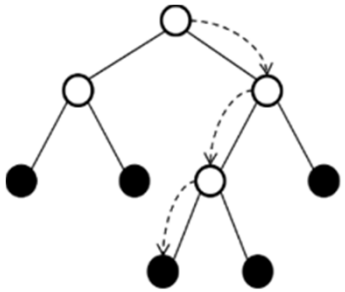




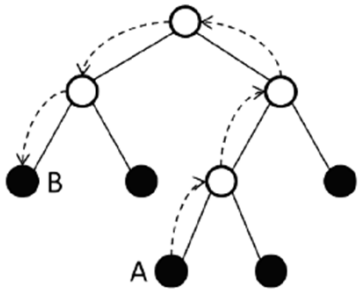


Il modello Evolving Tree (E-TREE)

Pakkanen, J., 2003. The Evolving Tree, a new kind of self-organizing neural network. Workshop on Self-Organizing Maps



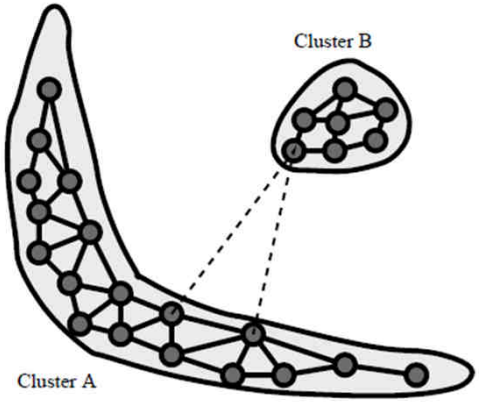
La ricerca del BMU di un nodo avviene a partire dalla radice e scendendo verso i nodi foglia



La vicinanza tra due nodi è definita dal numero di archi che separano due nodi. La distanza tra A e B è uguale a 5

Il modello Evolving SOM (E-SOM)

Deng D., Kabasov N., 2003. On-line pattern analysis by evolving self-organizing maps. Neurocomputing 51, Elsevier, 87-103

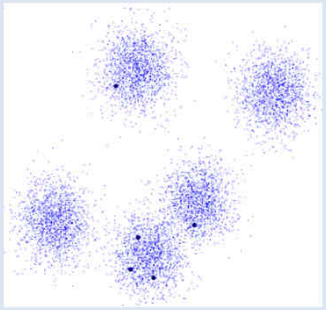


Le componenti connesse rivelano i cluster presenti nel dataset

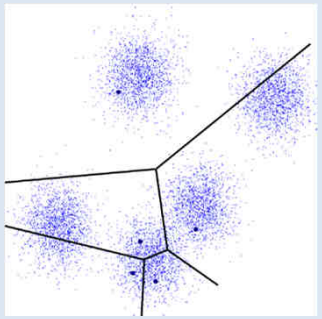
K-Means

Hartigan, J. A., Wong, M. A., 1979. "A K-means clustering algorithm". Applied Statistics, 28

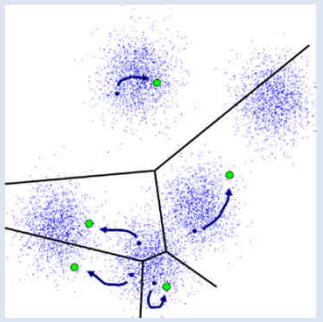
Partizionamento dello spazio in input in K partizioni, assegnando ogni input al centroide più vicino



Centroidi casuali



Partizionamento

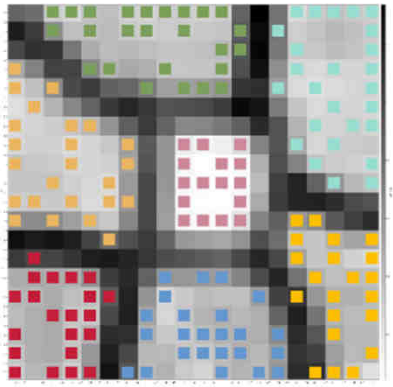


Riposizionamento centroidi

U-Matrix a componenti connesse (Umat-CC)

Hamel L., Brown C.W., 2011. Proceedings of the 2011 International Conference on Data Mining

Assegnazione di un nodo ad un cluster seguendo un percorso sulla mappa indicato dal nodo adiacente con gradiente più basso

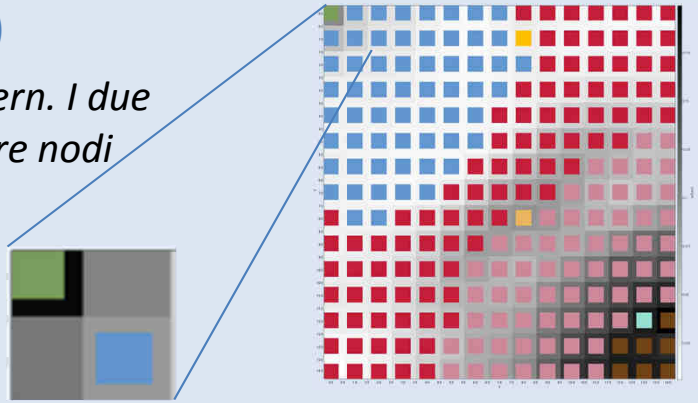


Sullo sfondo è presente la classica U-Matrix su scala di grigi. I colori sovrapposti indicano il riconoscimento automatico dei cluster

Two Winners Linkage (TWL)

Connessione dei due BMU di ogni pattern. I due nodi da connettere non devono essere nodi esterni sulla U-Matrix

*In verde è indicato un **nodo esterno***



Non connettendo i nodi esterni si riescono ad individuare le diverse zone della mappa e a riconoscere eventuali outliers

Tecnologie utilizzate

Software implementato da zero per rispettare i vincoli imposti dalla piattaforma DAMEWARE

Librerie utilizzate:

- **CFITSIO**

  Utilizzata per la conversione delle immagini astronomiche dal formato FITS ad ASCII

- **DevIL** (Developers Image Library)

  Utilizzata per la conversione delle immagini dai formati JPEG, GIF e PNG in ASCII, nonché per la creazione delle immagini fornite come output

- **STILTS** (Starlink Table Infrastructure Library Tool Set)

  Utilizzata per la conversione di file tabulari dai formati csv, votable e file FITS contenenti tabelle ad ASCII e per la creazione di istogrammi da fornire come output

Indice dei contenuti

1. Introduzione alle reti neurali: modello SOM
2. Evoluzioni del modello SOM
   - Evolving SOM ed Evolving Tree
   - Two-Stage Clustering
3. Tecnologie utilizzate
4. Introduzione ai test: indici di qualità
   - Test 1: Iris (confronto con algoritmo di clustering standard: K-Means)
   - Test 2: Chainlink
   - Test 3: Target
   - Test 4: M101 (immagine astronomica a banda singola)
5. Conclusioni

Indici di qualità

Errore di quantizzazione

$$QE = \frac{1}{N}\sum_{i=1}^{N}\|\overrightarrow{w_{BMU_i}} - \vec{x_i}\|$$

Similarità degli input assegnati al medesimo BMU

$\overrightarrow{w_{BMU_i}}$ = vettore dei pesi dell' i_{esimo} BMU
N = numero di pattern che compongono il dataset
$\vec{x_i}$ = i_{esimo} vettore in input rappresentato dal BMU considerato

Errore topografico

$$TE = \frac{1}{N}\sum_{i}^{N} u(\vec{w_i})$$

Dissimilarità degli input assegnati a BMU differenti

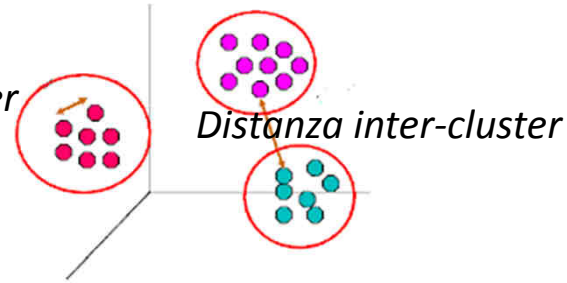
N = numero di pattern che compongono il dataset
$u(\vec{x_i}) = \begin{cases} 1, & \text{se il primo e il secondo BMU dell'} i_{esimo} \text{ pattern non sono adiacenti} \\ 0, & \text{altrimenti} \end{cases}$

Indice di Davies-Bouldin

$$DB = \frac{1}{k}\sum_{i=1}^{K}\max_{j,j\neq i}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}\right\}$$

Rapporto tra distanza intra-cluster ed inter-cluster

$S_{i,q} = \frac{1}{|C_i|}\sum_{\vec{x_i}\in C_i}\{|\vec{x_i} - \vec{z}|^q\}^{1/q}$, $i = 1 \dots K$ distribuzione interna dei cluster
$d_{ij,t} = |\vec{z_i} - \vec{z_j}|_t = \left\{\sum_{s=1}^{D}|z_{si} - z_{sj}|^t\right\}^{1/t}$ distanza tra due cluster



Indice di accuratezza

Rapporto tra cluster attesi ed individuati

$$ICA = \frac{|NC_c - NC_t|}{NC_c + NC_t}$$

Indice di completezza

Disgiunzione dei cluster

$$ICC = 1 - \frac{NC_d}{NC_t}$$

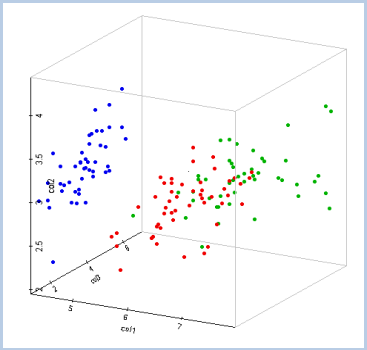
NC_t = numero di cluster teorici
NC_c = numero di cluster calcolati
NC_d = numero di cluster disgiunti

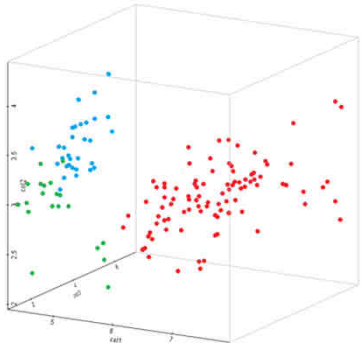
Indice dei contenuti

1. Introduzione alle reti neurali: modello SOM
2. Evoluzioni del modello SOM
   - Evolving SOM ed Evolving Tree
   - Two-Stage Clustering
3. Tecnologie utilizzate
4. Introduzione ai test: indici di qualità
   - Test 1: Iris (confronto con algoritmo di clustering standard: K-Means)
   - Test 2: Chainlink
   - Test 3: Target
   - Test 4: M101 (immagine astronomica a banda singola)
5. Conclusioni

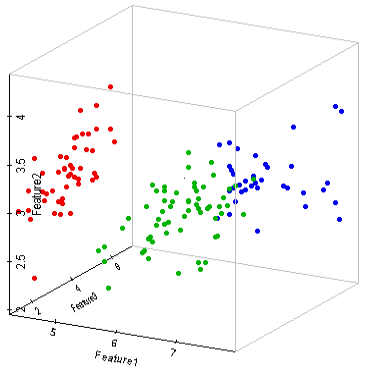
IRIS

Confronto tra
K-Means standard
e modelli di clustering
proposti

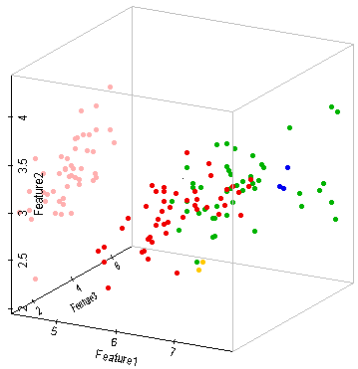




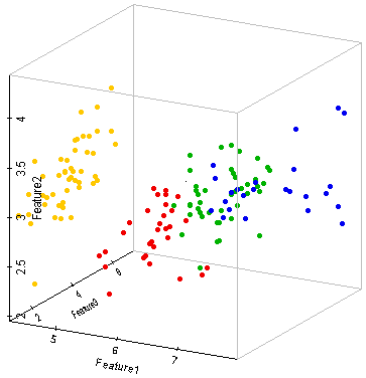
K-Means standard



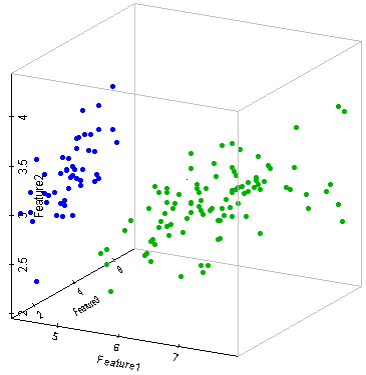
SOM + K-Means



SOM + TWL



SOM + Umat-CC



E-SOM

CLASSE	trovati	persi	%
Iris Setosa	50	0	100%
Iris Virginica	48	2	96%
Iris Versicolor	36	14	72%

Percentuale associazione
SOM + K-Means

CLASSE	trovati	persi	%
Iris Setosa	50	0	100%
Iris Virginica	45	5	90%
Iris Versicolor	47	3	94%

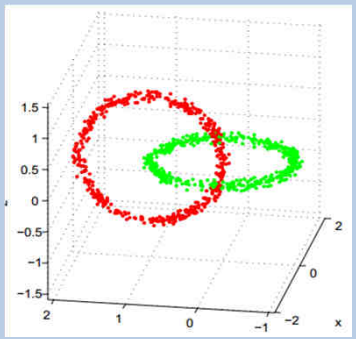
Percentuale associazione
SOM + TWL

Indice dei contenuti

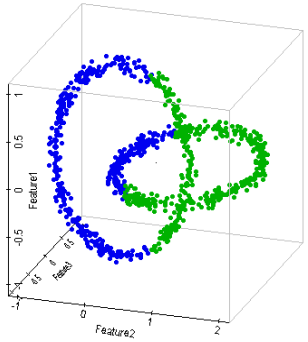
1. Introduzione alle reti neurali: modello SOM
2. Evoluzioni del modello SOM
   - Evolving SOM ed Evolving Tree
   - Two-Stage Clustering
3. Tecnologie utilizzate
4. Introduzione ai test: indici di qualità
   - Test 1: Iris (confronto con algoritmo di clustering standard: K-Means)
   - Test 2: Chainlink
   - Test 3: Target
   - Test 4: M101 (immagine astronomica a banda singola)
5. Conclusioni

Chainlink

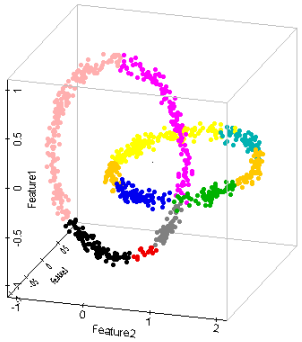
Confronto tra modelli di clustering proposti



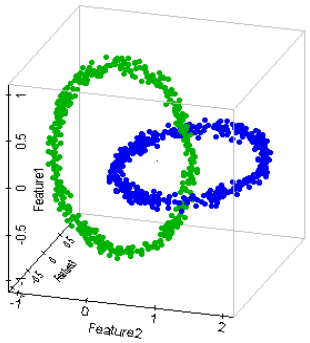
Dataset composto da due cluster con la caratteristica di essere non linearmente divisibili



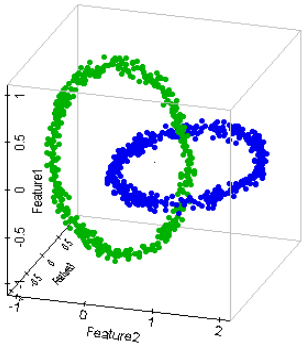
SOM + K-Means



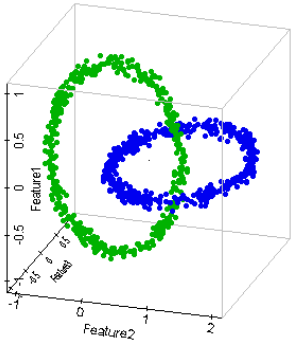
SOM + Umat-CC



SOM + TWL



E-SOM

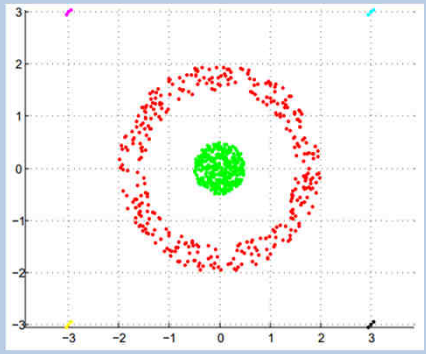


E-TREE

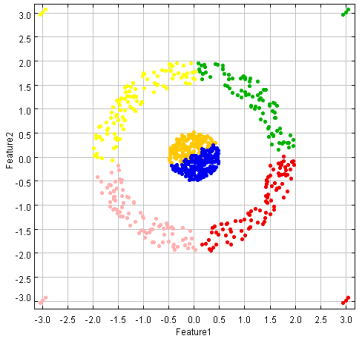
CONFRONTO STATISTICO SU PRESTAZIONI DI CLUSTERING (Chainlink)						
MODELLO	E-TREE	SOM	SOM+K-means	SOM+UmatCC	SOM+TWL	E-SOM
errore di quantizzazione	-	0.05	0.05	0.05	0.05	0.12
errore topografico	-	0.28	0.28	0.28	0.28	-
indice di Davies-Bouldin	-	-	1.15	0.68	2.02	1.99
Accuratezza	0		0	0.69	0	0
Completezza	0		1	0	0	0

Target

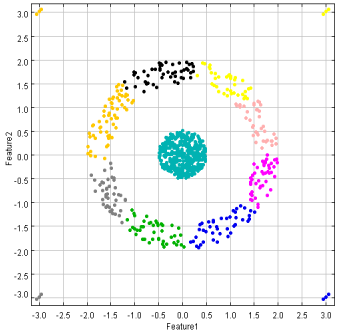
Confronto tra modelli di clustering proposti



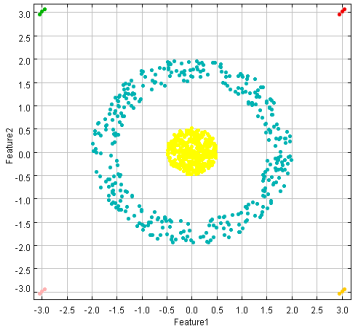
Dataset composto da cluster non linearmente divisibili e con quattro gruppi di outliers agli angoli dell'immagine



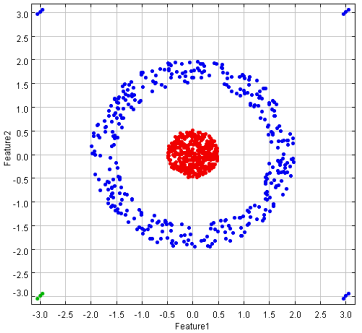
SOM + K-Means



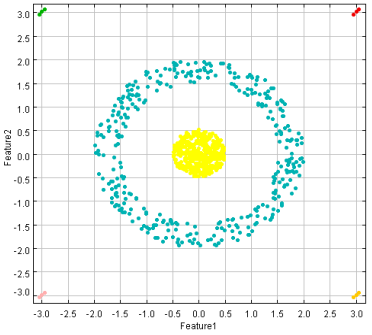
SOM + Umat-CC



SOM + TWL



E-SOM

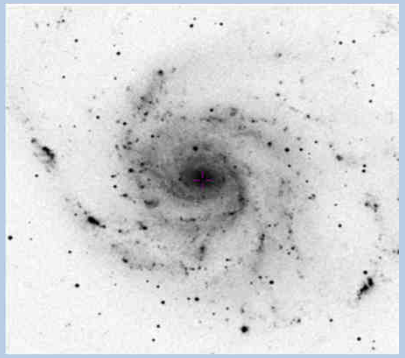


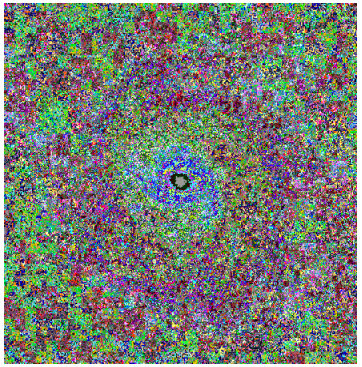
E-TREE

CONFRONTO STATISTICO SU PRESTAZIONI DI CLUSTERING (Target)						
MODELLO	E-TREE	SOM	SOM+K-means	SOM+UmatCC	SOM+TWL	E-SOM
errore di quantizzazione	-	0.07	0.07	0.07	0.07	0.09
errore topografico	-	0.07	0.07	0.07	0.07	-
indice di Davies-Bouldin	-	-	0.86	0.72	12.51	12.09
Accuratezza	0		0	0.2	0	0.33
Completezza	0		1	0.48	0	0.34

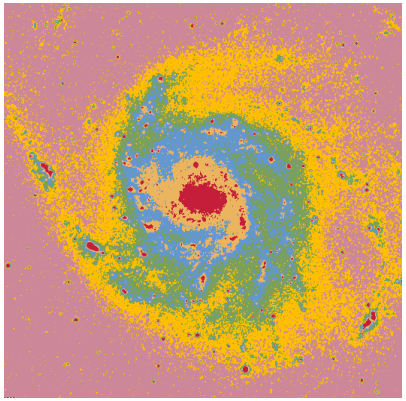
M101

Immagine astronomica monocromatica di una Galassia a spirale

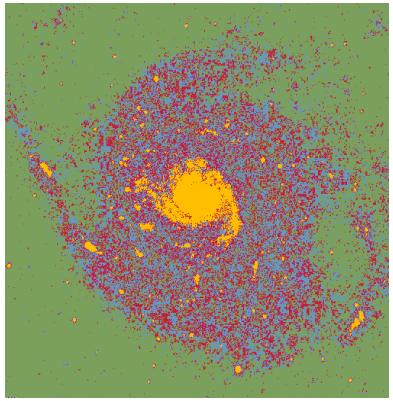




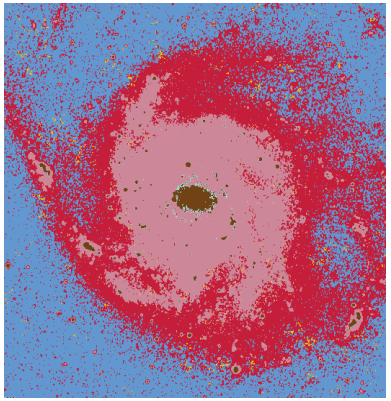
SOM (Single Stage)



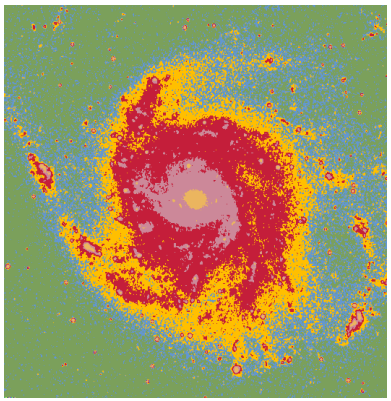
SOM + K-Means (k = 6)



SOM + Umat-CC



SOM + TWL



E-SOM

CONFRONTO STATISTICO SU PRESTAZIONI DI CLUSTERING (Img M101)					
MODELLO	SOM	SOM+K-means	SOM+UmatCC	SOM+TWL	E-SOM
errore di quantizzazione	0.001	0.001	0.001	0.001	0.03
errore topografico	0.90	0.90	0.90	0.90	-
indice di Davies-Bouldin	-	0.53	6.41	0.55	0.60

Confronto prestazioni tra metodi

	SOM + K-Means	SOM + Umat-CC	SOM + TWL	E-SOM	E-TREE
Dimensione dello strato di output	Ininfluente	Grande numero di nodi migliora i risultati	Grande numero di nodi aumenta la densità del clustering	Grande numero di nodi aumenta la densità del clustering	ininfluente
Outliers	Poco robusto	Poco robusto	Robusto	Robusto (dipendente da parametri)	Robusto
Dimensionalità dataset	Ininfluente	poco adatto su dataset 1D	Ininfluente	Ininfluente	Ininfluente
Dati linearmente non divisibili	Inefficace	Poco efficace	Efficace	Efficace	Efficace
Dipendenza dai parametri input	Alta	Media	Media	Alta	Alta

Conclusioni e sviluppi futuri

- Le diverse combinazioni di tecniche, nel clustering a due stadi, presentano caratteristiche tali da renderli adatti in diverse situazioni.
- Le evoluzioni incrementali del modello SOM possono essere usate in maniera ottimale sia per clustering che per classificazione
- Approfondimento del TWL, per migliorarne le prestazioni e la generalità d'uso