UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

Facoltà di Ingegneria
Corso di Studi in Ingegneria delle Telecomunicazioni

Tesi di Laurea in:

Analisi delle prestazioni di algoritmi di Machine Learning per la classificazione del traffico di rete

Relatore
Ch.mo prof. A. Pescapè
Correlatori
Ch.mo dott. M. Brescia
Ing. G. Aceto

Candidata
Luna Di Colandrea
N40/132

Anno Accademico
2012/2013

CONTESTO

Analisi e monitoraggio del traffico di rete :

Approcci di Machine Learning alla classificazione del traffico

CONTRIBUTO

- Studio di due innovativi algoritmi di Machine Learning
  - FMLPGA (Fast Multi Layer Perceptron with Genetic Algorithm)
  - GAME (Genetic Algorithm Modeling Experiment)
- Analisi e pre-processing delle tracce di traffico a livello pacchetto utilizzate
- Progetto e esecuzione di esperimenti sui due algoritmi
- Discussione dei risultati ottenuti

LA CLASSIFICAZIONE DEL TRAFFICO DI RETE

Con il termine **classificazione del traffico di rete** si intende l'associazione ad una sequenza di pacchetti scambiati tra due dispositivi e due rispettive porte a livello trasporto, della presunta applicazione che li ha generati

Importanza del tema

- sicurezza informatica
- servizi differenziati
- Studio di reti nei contesti reali
- Pianificazione e gestione di risorse di sistemi di TLC

Principali tecniche

Port-based	Inspection of packet payload	Flow-based	ML –based
Ispezione dei 16-bit del numero di porta del livello Trasporto	Controllo bit a bit del carico del pacchetto alla ricerca delle firme del protocollo	Controllo della sola intestazione, in cui sono contenuti i valori caratteristici	Imparano da dati empirici ad associare automaticamente ad un oggetto la classe di traffico
Pro: - Semplicità	**Pro:** - precisione	**Pro:** - no violazione della privacy	**Pro:** - generalizzazione, - indipendenza da modelli teorici o tecniche complesse di trattamento dati
Contro: - porte non standardizzate -Tecniche per aggirare filtri o firewall - NAT - FTP	**Contro:** - privacy - crittografia e offuscamento del protocollo - nuove applicazioni - grandi pacchetti	**Contro:** - Richiede la trasmissione di tutti i pacchetti del flusso	**Contro:** - Dipendenza dalla disponibilità di tracce per l'addestramento

→ algoritmi di ottimizzazione auto-adattativi

Applicazione Classificazione: associazione ad un campione della corrispondente etichetta (o classe)

Le fasi di classificazione :
-Fase di Training
-Fase di Testing
-Fase di valutazione

Paradigmi di apprendimento
-ML supervisionato
- ML non supervisionato

Esempi di metodologie di apprendimento supervisionato

Perceptron : 1 neurone di output e N neuroni di input(x)

$$y = H\left(\sum_{i=1}^{n} w_i x_i\right)$$

Multi-layer Perceptron : rete composta da uno o più livelli nascosti di neuroni, totalmente connessi tra i livelli di input e output

- MSE & "back propagation"

GLI ALGORITMI GENETICI

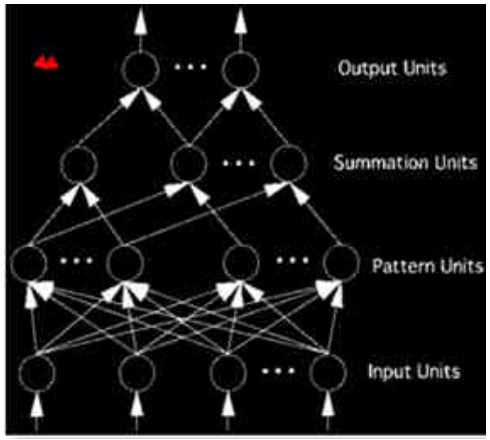
L'algoritmo genetico è un metodo di computazione ed ottimizzazione dei dati ispirato al processo evolutivo darwiniano .

Fasi di esecuzione:

1) Inizializzazione della popolazione
2) Selezione degli output
3) Confronto dei risultati
4) Evoluzione degli individui

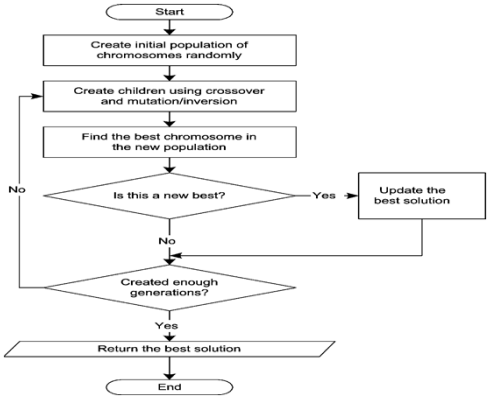
FMLPGA

❑ l'algoritmo genetico viene impiegato come sistema di addestramento del modello MLP al posto della Back Propagation



GAME

❑ E' un algoritmo genetico puro appositamente progettato per risolvere problemi di ottimizzazione in relazione a funzionalità di classificazione.



TRACCE DI TRAFFICO E IL DATASET

❑ **Le tracce di flusso** : file pcap contente 4.190.465 pacchetti , fornitoci dall'Università degli studi di Brescia[1] . Da questo file sono stati estratti due file : uno riguardante la Ground Truth e uno contenente i biflussi estratti .

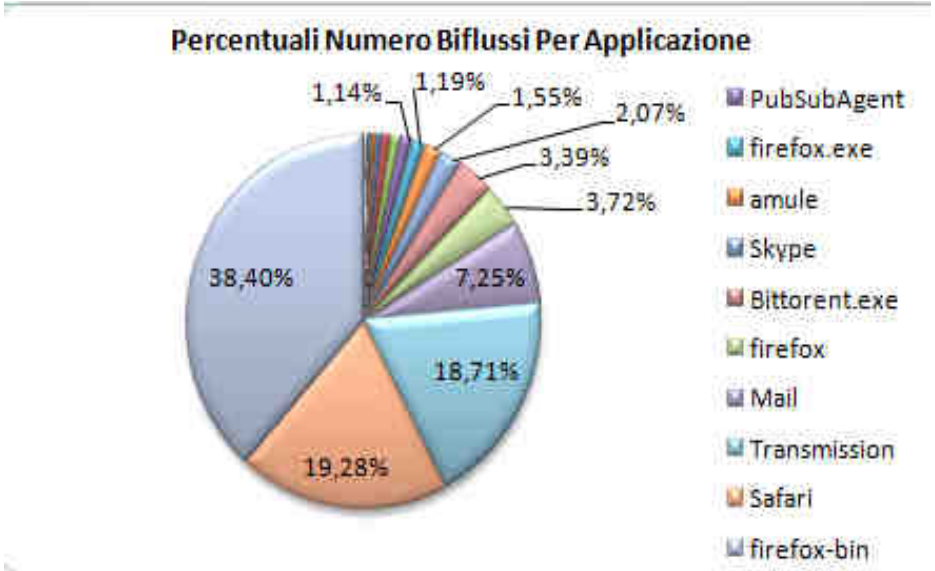
❑ **Biflusso**: insieme di pacchetti identificati dalla stessa quintupla:

- ip sorgente
- ip destinatario
- porta sorgente
- porta destinatario
- protocollo a livello trasporto TCP/UDP

in una comunicazione **bidirezionale.**

In questo elaborato di tesi i biflussi fanno riferimento al protocollo TCP.

❑ **Costituzione del dataset:** il dataset generale è costituito da 20251 biflussi composti da 4 caratteristiche e 31 applicazioni Targets .



❑ **Specificazione dei parametri di studio** :

- caratteristiche di input : time elapsed, byte, Up packet, Down packet
- classi Targets: applicazioni generatrici del traffico

1 http://www.ing.unibs.it/ntw/tools/traces

PRE-PROCESSING DEL DATASET

Dataset utilizzati : dalle 31 applicazioni contenute nel dataset originario sono state ricavate, per accorpamento, 6 classi principali : Browser Web, Skype, Mail, Traffico cifrato, P2P e Other. Eliminati i biflussi appartenenti alla classe Other sono stati creati 10 dataset biclasse contenenti tutti i possibili confronti delle 5 classi individuate .

Percentuali dei dataset : sono state usate sia le percentuali 70%(train)-30%(test) che 80%(train)-20%(test) per entrambi gli algoritmi.

Tipi di esperimenti : esperimenti Qualitativi (accuratezza) e Quantitativi (tempo)

GLI ESPERIMENTI

Scopo: capire quali classi di traffico sono più simili tra loro e quali invece risultano più facilmente classificabili.

Numero di prove : per ogni esperimento eseguito sono state effettuate tre prove identiche per un totale di 189 esperimenti per GAME e 129 esperimenti per FMLPGA.

Parametri variati:

- **GAME** : grado del polinomio 5, 8, 10 e funzione di selezione degli operatori genetici Roulette, Fitting , Ranking
- **FMLPGA:** funzione di selezione degli operatori genetici Roulette(V_1 e V_2), Fitting, Ranking

RISULTATI QUALITATIVI GAME

NETWORK TRAFFIC CLASSIFICATION - TEST BASED ON GAME												
70% train – 30% test				*these columns report classification accuracy percentages (average on two classes)*								
V	P	CASE	FUNC	Br-Ml	Br-PP	Br-Sk	Br-TC	Ml-PP	Ml-Sk	Ml-TC	PP-Sk	PP-TC	Sk-TC
CPU	5	TEST	ROULET.	73,75	63,68	57,82	68,02	79,43	75,54	73,62	62,13	63,43	73,17
			FITTING	74,05	63,29	57,52	68,02	79,35	75,97	73,04	62,59	64,4	71,82
			RANK.	74,05	63,2	57,82	67,12	79,48	76,25	72,75	62,36	63,11	72,09
	8	TEST	ROULET.	81,95	66,76	57,23	61,71	85,93	87,84	80	69,84	63,75	72,9
			FITTING	81,06	67,44	55,46	63,96	85,5	87,41	80,87	70,75	63,43	72,63
			RANK.	81,43	67,34	59	60,81	84,91	87,7	80,87	70,52	61,17	72,63
	10	TEST	ROULET.	81,58	66,47	59,29	64,86	85,25	86,84	80,58	68,03	64,08	73,98
			FITTING	81,8	67,44	58,7	66,22	85,46	88,41	83,19	68,25	62,78	74,25
			RANK.	82,48	66,57	58,7	65,32	85,38	87,41	80	69,55	63,43	72,63
		NUM PATTERNS		447	346	113	74	791	233	115	147	103	123

RISULTATI QUALITATIVI GAME

NETWORK TRAFFIC CLASSIFICATION - TEST BASED ON GAME													
70% train – 30% test				these columns report classification accuracy percentages (average on two classes)									
V	P	CASE	FUNC	B-Ml	Br-PP	Br-Sk	Br-TC	Ml-PP	Ml-Sk	Ml-TC	PP-Sk	PP-TC	Sk-TC
CPU	5	TEST	ROULET.	73,75	63,68	57,82	68,02	79,43	75,54	73,62	62,13	63,43	73,17
			FITTING	74,05	63,29	57,52	68,02	79,35	75,97	73,04	62,59	64,4	71,82
			RANK.	74,05	63,2	57,82	67,12	79,48	76,25	72,75	62,36	63,11	72,09
	8	TEST	ROULET.	81,95	66,76	57,23	61,71	85,93	87,84	80	69,84	63,75	72,9
			FITTING	81,06	67,44	55,46	63,96	85,5	87,41	80,87	70,75	63,43	72,63
			RANK.	81,43	67,34	59	60,81	[illegible]	87,7	[illegible]	70,52	[illegible]	72,63
	10	TEST	ROULET.	81,58	66,47	59,29	[illegible]	85,25	86,84	80,58	[illegible]	64,08	73,98
			FITTING	81,8	67,44	58,7	[illegible]	85,46	88,41	83,19	[illegible]	62,78	74,25
			RANK.	82,48	66,57	58,7	65,32	[illegible]	87,41	80	69,55	63,43	72,63
		NUM PATTERNS		447	346	113	74	791	233	115	147	103	123

	Ml-PP	Ml-Sk	Ml-TC
ROULET.	85,25	86,84	80,58
FITTING	85,46	88,41	83,19
RANK.	85,38	87,41	80
	791	233	115

Il dataset con i migliori risultati è Mail-Skype

Matrice di Confusione caso Mail–Skype

TRAINING	P(C1 Sk)	P(C2 Ml)
target class 1	219	47
target class 2	15	260

TEST	P(C1 Sk)	P(C2 Ml)
target class 1	103	18
target class 2	7	105

total classification	88,54%
class 1 classification	82,3308%
Class 2 classification	94,5455%

total classification	89,27%
class 1 classification	85,12%
Class 2 classification	93,75%

RISULTATI QUALITATIVI FMLPGA

NETWORK TRAFFIC CLASSIFICATION - TEST BASED ON FMLPGA												
80% train -- 20% test			*classification accuracy percentages (average on two classes)*									
V	CASE	FUNC	Br-Ml	Br-PP	Br-Sk	Br-TC	Ml-PP	Ml-Sk	Ml-TC	PP-Sk	PP-TC	Sk-TC
CPU	TEST	ROULV1	91,61	78,79	91,89	79,59	93,74	94,19	93,51	78,57	71,43	91,46
		ROULV2	91,95	81,82	90,54	75,51	93,36	93,55	92,21	86,73	87,14	92,68
		FITTING	89,6	71,43	91,89	73,47	90,89	90,32	93,51	79,59	72,86	92,68
		RANKING	89,93	79,65	91,89	81,63	91,46	93,55	92,21	85,71	81,43	91,46
	NUM PATTERNS		298	231	74	49	527	155	77	98	70	82
GPU	TEST	ROULV1	90,27	79,65	97,3	81,63	92,6	92,9	94,81	93,88	71,43	91,46
		ROULV2	91,61	81,82	91,89	85,71	91,84	94,84	93,51	83,67	68,57	92,68
		FITTING	90,27	80,95	90,54	79,59	92,03	94,84	93,51	92,86	70	92,68
		RANKING	91,28	80,82	93,15	84,1	91,45	92,74	92,98	87,62	70,32	90,14
	NUM PATTERNS		298	231	74	49	527	155	77	98	70	82

RISULTATI QUALITATIVI FMLPGA

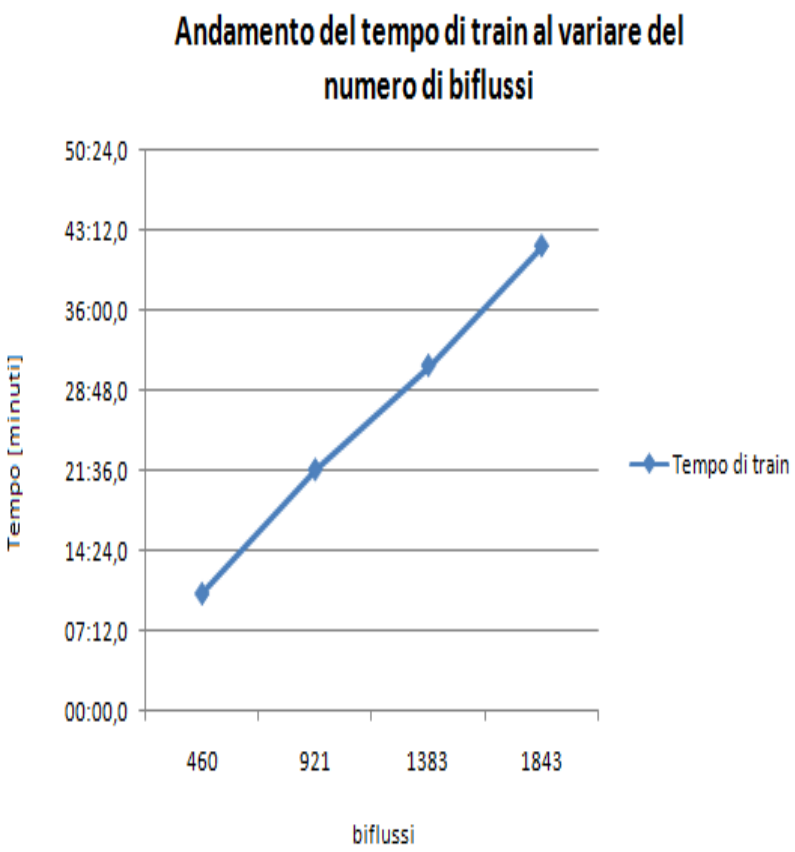
NETWORK TRAFFIC CLASSIFICATION - TEST BASED ON FMLPGA												
80% train -- 20% test			*classification accuracy percentages (average on two classes)*									
V	CASE	FUNC	Br-Ml	Br-PP	Br-Sk	Br-TC	Ml-PP	Ml-Sk	Ml-TC	PP-Sk	PP-TC	Sk-TC
CPU	TEST	ROULV1	91,61	78,79	91,89	79,59	93,74	94,19	93,51	78,57	71,43	91,46
		ROULV2	91,95	81,82	90,54	75,51	93,36	93,55	92,21	86,73	87,14	92,68
		FITTING	89,6	71,43	91,89	73,47	90,89	90,32	93,51	79,59	72,86	92,68
		RANKING	89,93	79,65	91,89	81,63	91,46	93,55	92,21	85,71	81,43	91,46
	NUM PATTERNS		298	231	74	49	527	155	77	98	70	82
GPU	TEST	ROULV1	90,27	79,65	97,3	81,63	92,6	92,9	94,81	[illegible]	71,43	91,46
		ROULV2	91,61	81,82	91,89	85[illegible]	91,84	94,84	93,51	83,67	68[illegible]7	92,68
		FITTING	90,27	80,95	90,54	7[illegible]5	92,03	94,84	93,51	92,86	7[illegible]	92,68
		RANKING	91,28	80,82	93,15	84,1[illegible]	91,4[illegible]	92,74	92,98	87,62	[illegible],32	90,14
	NUM PATTERNS		298	231	74	49	527	155	77	98	70	82

	Ml-PP	Ml-Sk	Ml-TC	PP-Sk
ROULV1	92,6	92,9	94,81	93,88
ROULV2	91,84	94,84	93,51	83,67
FITTING	92,03	94,84	93,51	92,86
RANKING	91,45	92,74	92,98	87,62
	527	155	77	98

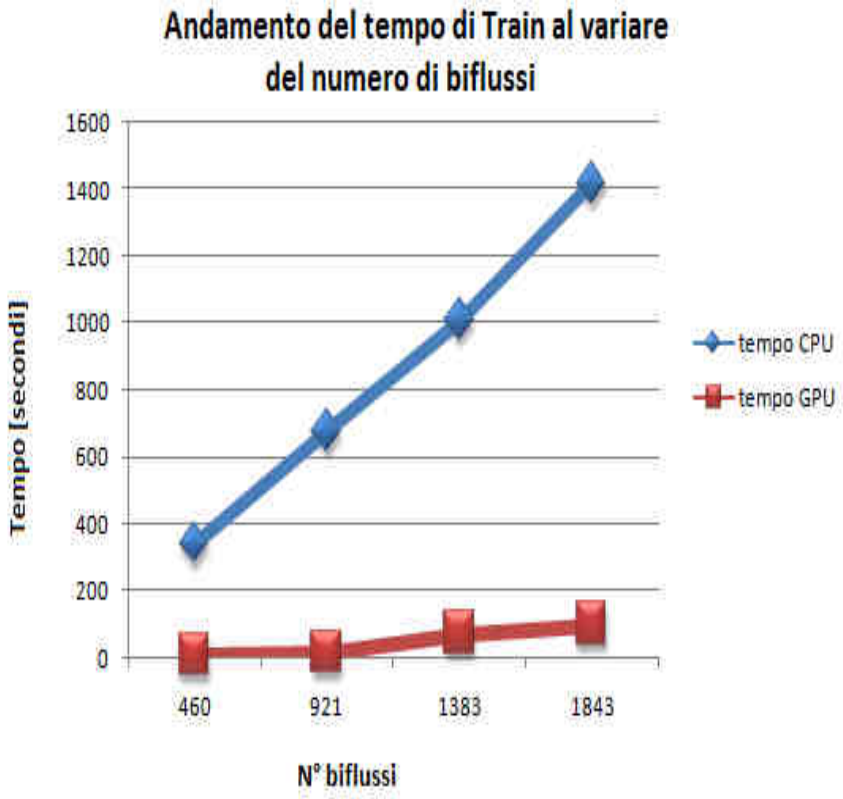
Le percentuali di classificazione risultano più omogenee rispetto a GAME

RISULTATI QUANTITATIVI

GAME



FMLPGA



Speed Up : 25x

CONCLUSIONI

In questo elaborato di tesi abbiamo:

- ❑ introdotto un pre-processing delle tracce di traffico per poter utilizzare gli algoritmi di classificazione considerati
- ❑ analizzato le prestazioni di due algoritmi di ML per la classificazione del traffico di rete

LAVORI FUTURI

- ❑ Esecuzione di esperimenti di tipo multiclasse
- ❑ Studio delle prestazioni di GAME su architetture GPU
- ❑ Esperimenti su altre tracce di traffico contenenti anche altre classi di traffico

Grazie per l'attenzione