Network Protocol Identification Ensemble with EA Optimization

[Extended Abstract]

R. G. Goss
University of Cape Town
Department of Computer Science
Cape Town, South Africa
ryan@goss.co.za

G. S. Nitschke
University of Cape Town
Department of Computer Science
Cape Town, South Africa
gnitschke@cs.uct.ac.za

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Concept Learning, Parameter Learning*

General Terms

Algorithms,Experimentation

Keywords

Evolutionary Algorithms, Artificial Neural Networks, Network Traffic Classification

1. EXTENDED ABSTRACT

In computer networks, the ability to correctly classify and control traffic flows is essential in order to manage network resources [9]. A number of works have focused on the identification of flow attributes, or discriminators, able to distinguish the underlying application protocol of a flow at an early stage of it's existence (see Bernaille *et al.* [1], Gargiulo *et al.* [3]).

In this study k-means is investigated for identifying distinct application protocols present within flow data sets generated using a select number of discriminators. Moreover, the k-means algorithm is used to identify and cluster similar flows. These clusters are subsequently used to train supervised *Artificial Neural Network* (ANN) classifiers in identifying future instances of the protocol. This paper first describes the identification and classification method, thereafter the utility of the proposed method is tested against live network traffic.

Implementation of the k-means algorithm is dependent on the number of clusters within the data set, k, known *a priori*. Most often, the value of k is chosen by running the clustering algorithm with different values of k, choosing the best using a predefined criterion [6]. In this work, an *Evolutionary Algorithm* (EA) is used to identify the most optimal value for k, as the data set may at any given time describe an arbitrary number of distinct application protocols (clusters). Each chromosome, or genotype, within the population represents an integer value for k, encoded as a binary string. The search space of the EA is configured as the lowest power of 2 such that $\sqrt{n}$ is divisible at least once, where n is the number of samples in the data set. The population of the first generation consists of genotypes with randomly generated values. At each generation, the fitness of each genotype is evaluated by way of a fitness function. This fitness function instantiates and executes an instance of the k-means algorithm, with the value of each genotype as k. Upon convergence, analysis of the clustered data set is performed using the *Silhouette Cluster Analysis* method, described by Rousseeuw [8]. The average silhouette, $s(i)$, of each cluster identified in the data set is a measure of how tightly grouped the members of the cluster are. The average silhouette for the clustered data set as a whole is therefore a measure of how successful the clustering process was. The fitness function scores each genotype with a value ranging from -1 through 1, with a value toward the latter signifying more optimal clustering and subsequently a fitter genotype. After the population has been evaluated, the process of *elitism* is performed whereby the highest scoring genotype is immediately transferred to the next generation's population. The remaining genotypes of the next generation are populated through a selective breeding process. Two genotypes from the current population are selected using *roulette selection* and bred using a *one-point crossover* method. *Bit string mutation* is applied during the breeding process, ensuring increased genetic diversity within the new population. A set number of generations are executed in this fashion, after which time the winning genotype and its associated data set declared the overall winner. Each cluster is then, in turn, used to train an ANN in the identification of future instances of the protocol, in accordance with Goss & Botha [4].

An experiment was conducted to ascertain the utility of classifying network traffic flows by applying the proposed method. Flow inspector software was used to record various discriminators from passing flows, in accordance with [4]. A total of 2488 flow samples were collected over a 1 hour period on a corporate wide area network (WAN) connection during a normal business day. The data set was clustered using a k-means approach with an EA determining the best value of k. The search space was set to 64, the smallest power of 2 in which $\sqrt{2488} = 49,88$ is divisible at least once. At each generation, the fitness of the population was determined using silhouette cluster evaluation. The EA determined the best value for k as 16, with an $s(i)$ value of 0.916821. This value was observed whilst scoring the population of the third generation. Of the 16 clusters identified, only those with a high number of samples, relative to the size of the data set,

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

were considered for training ANN classifiers. Training an ANN using limited samples could lead to an overly specific classifier (over-fitting), described by [2]. The observation of low sample counts within a cluster was likely attributed to the small observation window used to sample network traffic flows. After pruning, the clustered data set consisted of 11 clusters in total. The discriminator sets associated to each of these clusters was used to train an ANN classifier in the recognition of similar flows, as described in [5].

A sample of each protocol observed by Goss & Nitschke [5] was tested against the newly trained classifiers in order to evaluate the success of the proposed method. The highest scoring classifier and it's result along with the results obtained by [5] for each protocol are listed in table 1.

Data Set	Classifier ID	Certainty	Goss & Nitschke [5]
POP3	N/A	N/A	99.88%
SMTP	4	99.94%	99.86%
IMAP	13	99.86%	99.77%
HTTP	1	99.88%	69.60%
HTTPS	14	99.89%	99.35%
Soul-Seek	N/A	N/A	99.83%
Bittorrent	11	98.78%	99.70%

Table 1: Trained ANN accuracy

Table 1 shows an improvement in results for most of the protocols tested by [5]. Most noteworthy is the improved recognition of the HTTP protocol, which in [5] scored a mere 69.60%. These results demonstrate that by using k-means clustering with an EA selected k value, it is possible to achieve a higher degree of accuracy in clustering flow discriminator sets than when using a *Hierarchical Self Organizing Map* (HSOM), described in [5]. The EA furthermore demonstrated an optimal selection process for determination of k, identifying the best value within the first three generations. The use of an EA to identify the value of k is therefore more efficient than testing for the same sequentially.

Although the method described in this work achieved more accurate results than that of Goss & Nitschke [5], it isn't without it's limitations. K-means, like Gaussian Mixture EM clustering, is unable to adequately find non-linearly separable clusters within a data set. The Isodata type family of clustering to which k-means belongs, can only make convex clusters due to the nearest-neighbour clustering rule [7]. Furthermore, k-means does not have a notion of noise and therefore less than optimal clustering may be achieved in noisy data sets.

Future work will therefore investigate the application of the DBSCAN algorithm in a similar setting. The two parameters required by DBSCAN, namely the maximal neighbourhood radius, ϵ, and the minimal points per cluster, will be identified and evaluated using an EA. The DBSCAN algorithm is able to find arbitrary shaped clusters and also operate within a noisy data set. It is anticipated that these features will assist in the production of even more optimal clustering and, as a result, the development of more accurate classifiers.

2. REFERENCES

[1] L. Bernaille, R. Teixeira, and K. Salamatian. Early application identification. In *Proceedings of the 2006 ACM CoNEXT conference*, 2006.

[2] S. Cho and K. Cha. Evolution of neural network training set through addition of virtual samples. In *in Proc. 1996 IEEE Int. Conf. Evolutionary Computation,ICEC'96*, pages 685–688. IEEE Press, 1996.

[3] F. Gargiulo, L. Kuncheva, and C. Sansone. Network protocol verification by a classifier selection ensemble. In *Proceedings of MCS*, pages 314–323, 2009.

[4] R. Goss and R. Botha. Establishing discernible flow characteristics for accurate, real-time network protocol identification. In *Proceedings of the 2012 International Network Conference (INC2012)*, 2012.

[5] R. G. Goss and G. S. Nitschke. Automated network application classification: A competitive learning approach. In *In, Proceedings of the IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2013)*, 2013.

[6] A. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Lett.*, 2009.

[7] J. Lampinen and E. Oja. Clustering properties of hierachical self-organizing maps. *Mathematical Imaging and Vision*, 2:261–272, 1992.

[8] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and valudation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[9] R. Singh, H. Kumar, and R. Singla. Traffic analysis of campus network for classification of broadcast data. In *First International Conference on Intelligent Infrastructure the 47th Annual National Convention at COMPUTER SOCIETY of INDIA CSI*, 2013.