Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.187 – 194

A PEER TO PEER TRAFFIC IDENTIFICATION METHOD USING K-MEANS CLUSTERING, SVM & GENETIC ALGORITHM

Ruchika Aggarwal, Latika Gupta, Nanhay Singh

Aryabhatta College, University of Delhi, New Delhi, India

AIACTR, New Delhi

{Ruchika.aggarwal1989, Latikagup, Nsingh1973}@gmail.com

ABSTRACT: The utilization of shared (P2P) applications is developing significantly, which brings about a few major issues, for example, the system clog and traffic obstruction. Consequently P2P traffic identification is the most blazing theme of P2P traffic administration. Support vector machine (SVM) has points of interest with settling little examples for P2P characterization issues. However, the execution of SVM is basically reliant on its parameters. In this paper we propose Genetic Algorithm and K-Means with SVM to streamline the parameters of SVM and have been connected to P2P traffic identification. The curiosity of the proposed strategy is that it uses just the extent of parcels traded between IPs inside seconds. The recognized components of the proposed technique lie in that quick calculation, high identification precision, and asset sparing capacity. At last, experiment results demonstrate the satisfactory performance of the proposed method.

Keywords- P2P, SVM, Genetic Algorithm, K-Means Algorithm

I. INTRODUCTION:

Overview

A distributed (P2P) system is gathering of PCs, each of which goes about as a hub for sharing records inside the gathering. Rather than having a focal server to go about as a common drive, every PC goes about as the server for the records put away upon it. At the point when a P2P system is built up over the Web, a focal server can be utilized to file documents, or a

dispersed system can be set up where the sharing of records is part between every one of the clients in the system that are putting away a given record.

In the most fundamental sense, a distributed system is a straightforward system where every PC serves as a hub and a server for the records it only holds. These are the same as a home system or office organizes. Nonetheless, when P2P systems are built up over the web, the span of the system and the documents accessible enable colossal measures of information to be shared. Early P2P systems like Napster utilized customer programming and a focal server, while later systems like Kazaa and BitTorrent got rid of the focal server and part up sharing obligations between different hubs to free up data transmission. Distributed systems are generally connected with Web theft and unlawful record sharing.

The underlying utilization of P2P systems in business took after the organization in the mid-1980s of unsupported PCs. As opposed to the small centralized computers of the day, for example, the Versus framework from Wang Research centres Inc., which served up word preparing and different applications to imbecilic terminals from a focal PC and put away records on a focal hard drive, the then-new PCs had independent hard drives and implicit CPUs. The savvy boxes likewise had on board applications, which implied they could be conveyed to desktops and be helpful without an umbilical rope connecting them to a centralized server.

In its least difficult frame, a distributed (P2P) system is made when at least two PCs are associated and share assets without experiencing a different server PC. A P2P system can be an impromptu association—a few PCs associated by means of a Widespread Serial Transport to exchange documents. A P2P arrange likewise can be a perpetual framework that connections about six PCs in a little office over copper wires. Or, on the other hand a P2P system can be a system on a considerably more amazing scale in which exceptional conventions and applications set up direct connections among clients over the Web.

In a P2P arrange, the "companions" are PC frameworks which are associated with each other by means of the Web. Records can be shared specifically between frameworks on the system without the need of a focal server. As it were, every PC on a P2P organizes turns into a document server and additionally a customer.

The main necessities for a PC to join a distributed system are a Web association and P2P programming. Regular P2P programming programs incorporate Kazaa, Limewire, BearShare, Morpheus, and Procurement. These projects associate with a P2P system, for example, "Gnutella," which enables the PC to get to a large number of different frameworks on the system.

When associated with the system, P2P programming enables you to scan for records on other individuals' PCs. Then, different clients on the system can scan for documents on your PC, however regularly just inside a solitary envelope that you have assigned to share. While P2P organizing makes document sharing simple and advantageous, is additionally has prompted a ton of programming robbery and illicit music downloads. Hence, it is best to be erring on the side of caution and just download programming and music from true blue sites.

Distributed (P2P) is a decentralized interchanges display in which each gathering hosts similar abilities and either get-together can start a correspondence session. Not at all like the customer display, in which the customer makes an administration ask for and the server satisfies the demand, the P2P arrange demonstrate enables every hub to work as both a customer and server.

P2P frameworks can be utilized to give anonymized steering of system movement, monstrous parallel figuring situations, dispersed capacity and different capacities. Most P2P projects are centred on media sharing and P2P is along these lines frequently connected with programming and copyright infringement.

Normally, distributed applications enable clients to control numerous parameters of operation: what number of part associations with look for or permit at one time; whose frameworks to interface with or dodge; what administrations to offer; and what number of framework assets to dedicate to the system. Some essentially associate with some subset of dynamic hubs in the system with little client control, be that as it may.

K-Means Algorithm

K- Means Algorithm is most normal and prevalent bunching device that is generally utilized as a part of numerous applications and it falls under the apportioning calculations that points in building the different examples and assesses them by utilizing some model. With the given gathering of n information, k diverse bunches are shaped with each group having an one of a kind centroid (mean) and therefore the apportioning is made. The letter k depicts the quantity of bunches should have been made. At the point when number of n articles is to be assembled into k groups, K bunch focus is to be instated. Each question will be given to the nearest group focus and .the focal point of bunch is refreshed each time until condition of no change happens in the each group. The components in each group will be in close contact with centroid of that specific bunch and will be distinctive to the components having a place with different bunches.

The aggregate of the disparities between the point and the centroid communicated by particular separation is utilized as the goal work. Add up to intra-group difference depicts the aggregate of the squares of the mistake between the point and separate centroids.

GENETIC ALGORITHM

Hereditary Calculations (GAs) are versatile heuristic inquiry calculation in view of the transformative thoughts of normal determination and hereditary qualities. Accordingly they speak to a shrewd misuse of an arbitrary pursuit used to tackle enhancement issues. Albeit randomized, GAs are in no way, shape or form arbitrary, rather they misuse verifiable data to coordinate the hunt into the district of better execution inside the inquiry space. The essential strategies of the GAs are intended to recreate forms in common frameworks important for advancement; particularly those take after the standards initially set around Charles Darwin of "survival of the fittest."

GAs depends on a similarity with the hereditary structure and conduct of chromosomes inside a populace of people utilizing the accompanying establishments:

- Individuals in a populace vie for assets and mates.
- Those people best in every "opposition" will create more posterity than those people that perform ineffectively.
- Genes from `good' people spread all through the populace so that two great guardians will now and again create
- Suited to their condition.

Genetic algorithm

1. Randomly introduce population (t)
2. Determine wellness of population (t)
3. Repeat
    - select guardians from population(t)
    - perform hybrid on guardians making population(t+1)
    - perform change of population(t+1)
    - determine wellness of population(t+1)
    - until best individual is sufficient

SUPPORT VECTOR MACHINE:

"Support Vector Machine" (SVM) is an administered machine learning calculation which can be utilized for both order and relapse challenges. Be that as it may, it is generally utilized as a part of arrangement issues. In this calculation, we plot every information thing as a point in n-dimensional space (where n is number of components you have) with the estimation of each element being the estimation of a specific arrange. At that point, we perform grouping by finding the hyper-plane that separate the two classes exceptionally well (take a gander at the underneath preview). Support Vectors are basically the co-ordinates of individual perception. Bolster Vector Machine is a wilderness which best isolates the two classes (hyper-plane/line).

Support vector machines (SVMs) are an arrangement of directed learning strategies utilized for characterization, relapse and anomaly's identification.

Support Vector Machines depend on the idea of choice planes that characterize choice limits. A choice plane is one that isolates between arrangements of items having diverse class participations. A schematic case is appeared in the representation beneath. In this case, the items have a place either with class GREEN or RED. The isolating line characterizes a limit on the correct side of which all articles are GREEN and to one side of which all items are

RED. Any new protest (white hover) tumbling to the privilege is marked, i.e., arranged, as GREEN (or named RED should it tumble to one side of the isolating line).

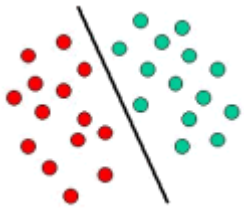

Fig: 1.1 Grouping of articles

Support Vector Machine (SVM) is basically a more tasteful strategy that performs order undertakings by developing hyperplanes in a multidimensional space that isolates instances of various class marks. SVM underpins both relapse and order errands and can deal with various persistent and straight out factors. For downright factors a spurious variable is made with case values as either 0 or 1. Therefore, an all-out ward variable comprising of three levels, say (A, B, C), is spoken to by an arrangement of three sham factors:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

To build an ideal hyperplane, SVM utilizes an iterative preparing calculation, which is utilized to limit a blunder work. As indicated by the type of the blunder work, SVM models can be characterized into four unmistakable gatherings:

- Classification SVM Sort 1 (otherwise called C-SVM characterization)
- Classification SVM Sort 2 (otherwise called nu-SVM characterization)
- Regression SVM Sort 1 (otherwise called epsilon-SVM relapse)

II. PROPOSED METHODOLOGY:

For P2P movement recognizable proof issue we proposed a joined approach utilizing unsupervised machine learning calculation K implies bunching for classes information in view of components. Support vector machines (SVM) are a standout amongst the most generally utilized machine learning techniques for arrangement and relapse issues of little specimens. Truth be told, the execution of SVM is to a great extent subject to its parameters determination. In the strategy of grouping by SVM, and high measurements, this has an extensive variety of uses, for example, picture order, confront discovery, content arrangement. SVM has a brilliant capacity to tackle the order issues for 2 classes. The principle reason for P2P activity recognizable proof is to precisely order two classifications: P2P and non-P2P movement. For this reason K-Means and SVM both will give more precise outcome. Hereditary calculation is a sort of reference organic common choice and normal hereditary system of the irregular hunt calculation; it is reasonable for taking care of complex issues which are hard to tackle by customary inquiry calculations. GA begins from the underlying irregular arrangement of arbitrary era; it creates new arrangements by a specific determination, hybrid and change operation well-ordered emphasis.

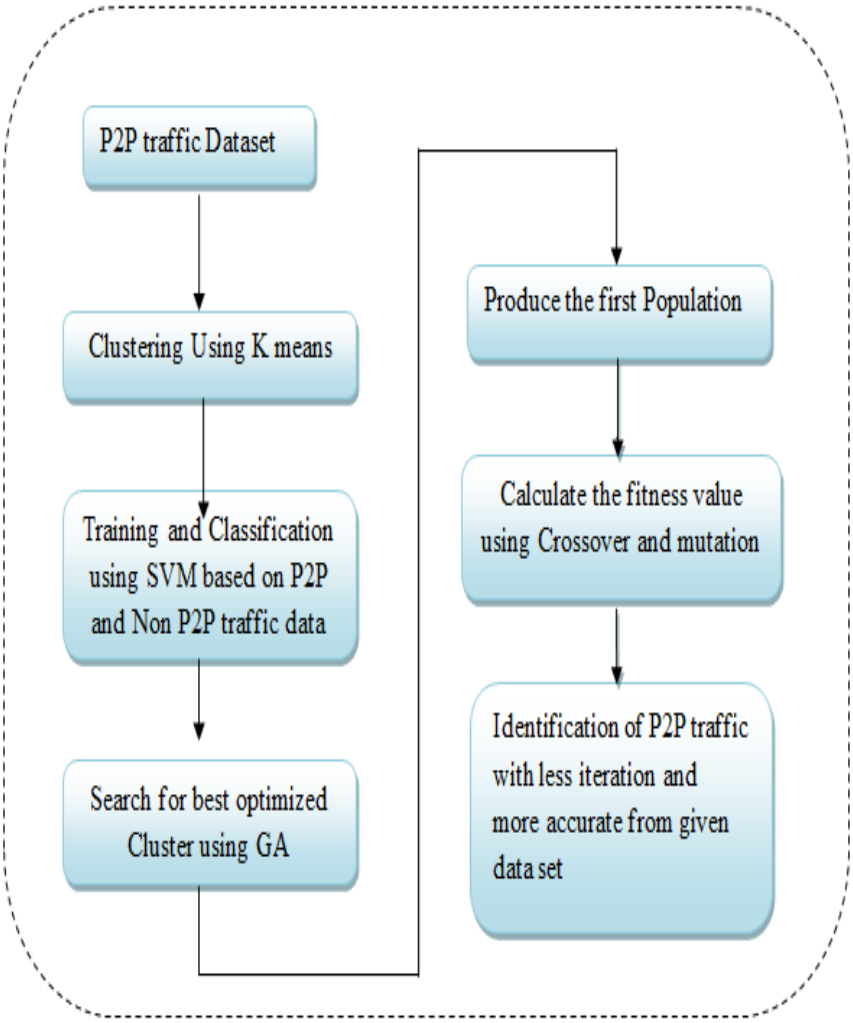

Fig: 1.2 Flowchart of Proposed Work

III. Result Analysis / Implementation:

The following Figure shows the response of energy consumption vs. transmission power traffic scenarios,

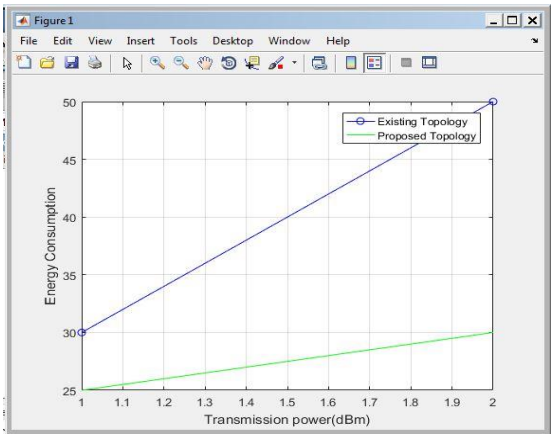

Fig: 1.3 Graph of energy consumption vs. transmission power

Network Traffic Information

We will give the preparation and stacking of datasets by applying characterization utilizing support vector machine (SVM) method. The following are the perceptions for stacking the informational index and preparing the informational collection in MATLAB apparatus.

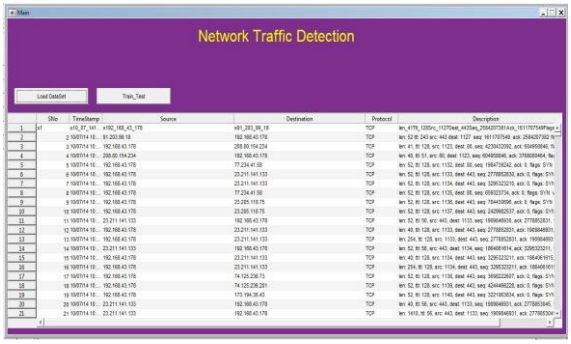

Fig: 1.4 shows the loading of dataset provided by SVM.

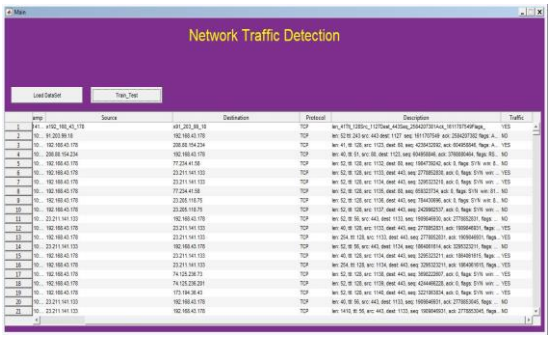

Fig: 1.5 Show the training of dataset provided by SVM.

The observations obtained by implementing simulation model for the traffic scenarios is provided in Table 1.1.The results are based on these observations.

Table 1.1: Observations for Varying Number of Node

Methodology	**Output**
Existing Methodology (Genetic Algorithm)	82.70 %
Proposed Methodology (K-Means, SVM and Genetic Algorithm)	87.37 %

IV. CONCLUSION:

The conclusions introduced in this exposition of system activity distinguishing proof gives indispensable favourable circumstances to IP arrange building, organization and control and other key spaces. Current acclaimed strategies, for instance, port-based and payload-based, have exhibited a couple inconveniences, and the machine learning based procedure is a potential one. The activity is requested by the payload-self-governing truthful characters. This paper exhibits the differing levels in system activity examination and the huge data in machine learning space, looking at the issues of port-based and payload-based techniques in

movement portrayal. Considering the need of the machine learning-based system, we attempt diverse things with K-means, SVM and GA to survey the productivity and execution. The trial happens on activity datasets pass on that the precision gained by our technique is progressed.

In this manner, all in all, the execution of P2P system activity is enhanced in productive way and with more precise outcomes.

REFERENCES

[01] Jie Cao, Zhiyi Fang, Dan Zhang, and Guannan Qu, "Network Traffic Classification Using Feature Selection and Parameter Optimization", Journal of Communications Vol. 10, No. 10, October 2015, doi:10.12720/jcm.v.n.p-p doi:10.12720/jcm.10.10.828-835.

[02] Prof S. R. Patil, Suraj Sanjay Dangat, "Identifying Peer-to-Peer Traffic Based on Traffic Characteristics", Recent Advances in Computer Science, ISBN: 978-1-61804-320-7.

[03] Satoshi Ohzahata, Yoichi Hagiwara, Matsuaki Terada, and Konosuke Kawashima, "A Traffic Identification Method and Evaluations for a Pure P2P Application".

[04] Joseph Stephen Bassi, Loo Hui Ru, Khammas, Muhammad, Nadzir Marsono, "Online Peer-To-Peer Traffic Identification Based On Complex Events Processing Of Traffic Event Signatures", Jurnal Teknologi (Sciences & Engineering) 78:7 (2016) 9–16, eISSN 2180–3722.

[05] Jinghua Yan, Zhigang Wu, Hao Luo, Shuzhuang Zhang, "P2P Traffic Identification Based on Host and Flow Behaviour Characteristics", CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No 3, Sofia • 2013, ISSN: 1314-4081, DOI: 10.2478/cait-2013-0026.

[06] Marcell Perényi, Trang Dinh Dang, András Gefferth and Sándor Molnár, "Identification and Analysis of Peer-to-Peer Traffic", Proceedings of 5th International IFIP-TC6 Networking Conference, Coimbra, Portugal, May, 2006. © 2006 IFIP.

