Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)
© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/APC220012

Distribution Diversity Method of Feature Optimization (DDMFO) to Defend the Intrusion Practices on IoT Networks

Bhargavi Mopuru[a,1] and Yellamma Pachipala[b]
[a]*Research Scholar, Department of CSE,*
[b]*Professor, Department of CSE,*
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P, India

Abstract. The new forms of networks labeled IoT are relatively new and which become buzz in this decade. The network architecture lets any smart device loosely connect to the Internet under internet protocol. However, the other dimension of this network facilitates intruders to access the network with no critical efforts. The context of intrusions has been delineated as intrusion practices of other devices connected to an IoT network that are connected to external networks through a gateway. Vice versa, the compromised IoT network intends to communicate with external devices or networks to perform intrusion practices. In this regard, intrusion detection through machine learning demands significant feature selection and optimization techniques. This manuscript endeavored to demonstrate the scope distribution diversity assessment methods of traditional statistical practices toward feature selection and optimization in this regard, the contribution "Distribution Diversity Method of Feature Optimization (DDMFO) to Protect Intrusion Practices on IoT Networks" of this paper uses the Dice Similarity Coefficient procedure to pick the optimum characteristics for the training of the classifier. The classifier that has been adopted in this contribution is Naïve Bayes, trained by the features selected by the proposal. The experimental research concludes the significance of the taxonomy, which demonstrates substantial accuracy and minimal false alarm.

Keywords: IoT, Feature Optimization, similarity coefficient, Distribution Diversity, Class Label Assessment.

1. Introduction

It is very clear that the Internet is widely available. The levels of its efficiency are high and, at the same time, versatile. Whereas the Internet has offered comfort in the lives and work of individuals, some of the main challenges presently being experienced include ensuring that the data pooled inside the network is highly secure. At the same time, the security of the equipment is of great concern to individuals.

The number of hackers in the world has been rising. It is also worth pointing out that network security problems have made the United States of America lose over 10

[1] Bhargavi Mopuru, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P, India; E-mail: bhargaviphd83@gmail.com.

billion U.S. dollars [1]. Other reports point out that theft of information has significantly risen at the rate of 250 percent, and 98 percent of well-known firms have in the past experienced attacks on their
networks. As a result, much emphasis has been put on the way in which the network can be made to be highly secure. The industry has strived to ensure that there are various kinds of measures in place which are aimed at enhancing the level of security of the networks. The academia has also strived to come up with various strategies which can be implemented to ensure that safety issues are handled in the right way. The industry has also not been left behind, as they have strived to invent various kinds of technologies that are aimed at helping in the detection of network attacks [2], [3].

IoT is generally a concept that was developed in the past years. Practically, it has been used in sensors, processors as well as in wireless communication modules embedded. It has also been used widely in railways, power grids, tunnels, bridges, buildings, and highways, as well as in other kinds of objects that are mutually connected. In the year 1999, MIT developed an automatic identification technology center, which was involved in the conception of the IoT concept founded on RFID. They developed the model of electronic product code. EPC system is capable of tracking goods in real-time. On the same note, they can optimize the entire supply chain to ensure that the users are supported. This has played a major role in promoting the rapid development of automatic identification technology. At the same time, it has been in a position to result in significant improvements in consumers' quality of life [4]. Intrusion behavior includes several things, including attempts to disrupt confidentiality, integrity, and access to the targeted resources. The security defense mechanisms, like authentication as well as encryption, are generally passive. Regardless of how the updates are done, they are often exposed to various kinds of attacks.

Detection of attacks is a new security technology that just erupted in the recent past. It can be of great significance when it comes to detecting network security attacks. With it, it is possible for people to adopt measures that are aimed at ensuring that changes are made to prevent intrusions. It also compensates for the challenges, which are associated with conventional security defense technology [5].

Denning developed a model for detecting intrusion [6]. He indicated that the detection of attacks ought to be founded on the collection of network packet information and a thorough analysis of the information which has been collected. At the same time, he indicated that the ability of attacks to take place should be determined, and the management ought to be informed in time for them to be in a good position to develop various kinds of strategies in order to deal with the attacks. As a result, a good system for detecting intrusions ought to include at least the requisite functions like gathering information regarding attacks, analyzing the given information, detecting the attacks in time, and alerting the management for them to be in a position to develop strategies to deal effectively with the attacks.

2. Related Work

IoT is generally exposed to numerous threats. As a result, there is a need to make sure that the technologies that are supposed to be used for detecting intrusions should not be compacted, should provide real-time information, and should be very accurate. This section offers a description of the most recent technologies which are being used for

detecting intrusions in the area of IoT.The models that are presently used for detecting attacks and for preventing attacks employ various kinds of statistical approaches [7] like the Hidden Markov Model [7], Bayes theory [8], cluster analysis [9], signal processing [10] as well as distance measuring [11] for the detection of activities, The techniques used for anomaly detection is divided into supervised learning, as well as unsupervised learning [12]. In the supervised anomaly detection techniques, a system's normal behavior or the networks is generally constructed through the use of datasets, which are labeled [13]. Unsupervised models work on the assumption that the normal node behaviors are generally more frequent and, as a result, this model is developed based on this viewpoint; therefore, there is no need for training data [14].

The work [15] suggested unsupervised NIDS founded on the clustering of subspace. He pointed out that their technique has a better performance against unknown intrusions. The work [16] developed a feature section filtering technique that uses PCA and FDR for filtering any kind of noise.

On the other hand, it has a very high false-positive rate. The work [17] suggested an unsupervised framework that was founded on the "Optimum-path forest algorithm" as well as the "K-Means clustering model". These context models malicious as well as typical behavior of the networks.

The work [18] suggested a two-level method of detecting attacks that commences with the identification of abuse and thereafter adopts the KNN algorithm to reduce false alarms.

The work [19] suggested a multi-classification intrusion technique that consists of support vector machines as well as a BRICH hierarchical clustering model for the extraction of important characteristics from the KDD99 dataset. This model has a very high rate of detection for DoS, as well as for Probe attacks. However, it is generally not highly effective against R2L as well as U2R intrusions.

The work [20] suggested a model for detecting DoS, which utilizes "multivariate correlation analysis (MCA)"for improving the characterization's accuracy of network traffic. The work [21] developed a two-layer classification technique for the detection of U2R as well as R2L intrusions. It has a low computational complexity because it has an optimized feature reduction. The work [22] suggested "an ensemble-based multi-filter feature selection technique" for the detection of distributed DoS intrusions within the cloud environments through the use of four filter techniques for the achievement of optimum selection above. "NSL-KDD dataset". The work [23] suggested intrusion taxonomy aimed at cloud services. He proposed a cloud-based system for detecting intrusions.

The work [24] suggests a communal data-based intrusion detection system that chooses optimal attributes for classification using a "feature selection algorithm." The technique was analyzed through the use of Kyoto 2006+, NSL-KDD as well as KDD Cup 99).

Industrial control systems make use of the systems for attack detection for them to be in a position to manage their security issues [25] better. The work [26] suggested a systematic as well as an automated technique for building "a hybrid intrusion detection system" that studies sequential state-based stipulations aimed at "electric power systems" to be in a position to precisely distinguish between ordinary control operations, disturbances, as well as cyber-attacks. The work [27] suggested filtering intrusions from actual errors, using a multi-method driven intrusion detection system and industrial anomaly oriented on the "Hidden Markov Model."

One of the main issues, which has significantly reduced the adoption of IoT devices, is mainly security issues [28]. The work [29] illustrated that a wide array of tools are capable of helping in the mitigation of cyber threats, which mainly target IoT systems. The work [30] proposed various leveled validation architectures for the arrangement of mysterious information transmission inside the IoT Networks. The work [31] Suggests the significance of ghost attacks on ZigBee oriented on IoT devices. The contribution [32] suggested an "autonomic model-driven cyber security management" technique for the IoT systems that may be employed for estimating, detecting, and responding to cyber-attacks with minimal or with no human interference. The work [33] suggested a scheme for preventing insider intrusion within the IoT networks through crosschecking the transformation of data of all IoT nodes.

3. Feature Optimization by Distribution Diversity

The method portrayed is a machine learning approach that functions in a sequence of learning and detection phases. The proposal's objective is to defend against intrusion practices that are switching by the external networks linked to the target IoT network. The learning phase of the proposed method uses the given records of the network transactions labeled either as positive or negative, which indicates the prone to intrusion or not in respective order. Further, the learning phase applies set theory. It identifies all possible unique subsets of the attributes, which represent the values in the given network transactions that fall in either of the class labels, positive or negative. Further, these subsets are sorted in ascending order of their size. Afterward, for each of these subsets, the values projected in different network transactions, both class labels are collected as a set such that each entry of this set is the pattern of values representing attributes of the corresponding subset. Later, the learning phase verifies the significance of these patterns in the records of both labels. This is done by the statistical method that assesses the distance between the values obtained for the set of attributes from the records named as positive or negative. Suppose the distance between the corresponding values obtained from the records labeled as positive and the pattern of values obtained from the records labeled as negative is observed as more than the given distance threshold. In that case, the corresponding pattern of attributes is identified as the optimal feature to train the classifier. In regard to measuring the distance, we opted for the statistical method called dice similarity coefficient.

A. The Features

The features used by the training phase of the proposed method DDMFO are the pattern of values projecting the pattern of attributes. In this regard, the phase that determines the features is as follows.

Let the given set NT of network transactions that are labeled either as positive or negative, which represents either prone to intrusion or not in respective order. Find the attributes as a set A, which represent the values of each record $\{r \exists r \in NT\}$ of the set NT.

Find all possible unique subsets of the attributes listed in the set A, which are listed as a set AS, such that each entry of the set AS is a unique subset $\{s \exists s \in AS \wedge s \subset A\}$ of the set A. Let the map F which is having a set VS mapped by a subset $\{s \exists s \in AS \wedge s \subset A\}$, and each entry $\{e \exists e \in VS \wedge e \in r\}$ of the set VS is the set of values projected in each record $\{r \exists r \in NT\}$ for the attributes of the corresponding subset s. Hence, an entry $\{s \to VS\}$ in the map F is the key s-value VS pair; here, the value VS is a set, and each entry e of this set VS is the pattern of values representing the attributes found in s the key in the corresponding sequence of attributes.

Further, it performs the following to enable the dice similarity coefficient can identify the optimal features from the map F

1. For each entry of the map F having a subset s of the attributes A as key,

Begin

a. let the set VS that mapped to key s,

b. List unique entries of the set VS as set UVS

c. Add the subset s and set UVS as key and value pair to the map UF

End

2. Let partition the records given as input to the training phase in to sets NT^+, NT^- such that these sets represent the records labeled as positive and negative in respective order.

3. For each subset $\{s \exists s \subset A\}$

Begin

a. For each set $\{UVS \exists \{s \to UVS\} \in UF\}$ that mapped to set s as key in a map F

b. For each record $\{r^+ \exists r^+ \in NT^+\}$ of the set NT^+ Begin

c. Move the index i of an entry $\{e \exists e \in UVS \wedge e \in r^+\}$ that exists in both set UVS s and record r^+ to the set V_s^+. This is the index of “pattern of values" in the set UVS representing the attributes of the set s in the record r^+

End

4. For each record $\{r^- \exists r^- \in NT^-\}$ of the set NT^-

Begin

a. Move the index i of an entry $\{e \exists e \in UVS \wedge e \in r^-\}$ that exists in both in set UVS and record r^- to the set V_s^-. This is the index of "pattern of values" in the set UVS representing the attributes of the set s in record r^-

End

B. Dice similarity coefficient

This model version is identified as a powerful method to find whether diversified or not for the given two vectors. According to previous statistical values, this method is optimal for two different values from the same distribution [34]. This method is adapted for the training set, which consists of positives and negatives and produces optimal features. The diversity of values in given two vectors v_1, v_2 denoted by Dice Similarity Coefficient dsc is estimated by using the following equation.

$$dsc = \frac{2^* | v_1 \bigcap v_2 |}{| v_1 | + | v_2 |}$$

In the equation above $| v_1 |, | v_2 |$ denotes the degree of the given vectors v_1, v_2, and the notation $| v_1 \bigcap v_2 |$ denotes the degree of the intersecting values of the given vectors v_1, v_2.

The dsc analysis that the two vectors are different if it is less than the given DSC threshold $dsct$ (usually $0.7 \le dsct < 1$).

C. Optimal Feature Selection

The features selected from the network transactions given as input to the training phase are listed as the sets $\{V_s^+ \exists s \subset A\}, and \{V_s^- \exists s \subset A\}$ for each subset of attributes. The dice similarity coefficient is used further to identify the distribution diversity between the respective set $\{V_s^+ \exists s \subset A\}, and \{V_s^- \exists s \subset A\}$, and if diversity is observed, then the subset s represented by the values from positive and negative records of the training set is considered optimal to classify the unlabelled network transactions.

D. Class Label Assessment

Each record is being processed for a given set of test records, t which extracts the pattern of values for all possible subsets of exists in the set AS. Further, the probability of both class labels for each corresponding pattern will be estimated. Then the fitness of the given test record towards the positive label will be estimated, which is the absolute difference of the average of the probabilities identified for all of the corresponding patterns toward the positive label and their deviation error. Similarly, the fitness of the given test record towards a negative label is also being estimated. This is the absolute difference of the average of the probabilities identified for all of the corresponding patterns toward negative label and their deviation error. Further, the test record will be labeled as positive if the fitness of the corresponding record t toward the positive label is greater than the fitness of the corresponding record t toward the negative label. If not, the fitness of the record towards the negative label is greater than the fitness of the record toward the positive label, and then the record will be labeled as unfavorable. In another case, if both fitness values respective to positive and negative labels are approximately equal, then the record will not be labeled and recommended for administrative decision.

4. Experimental Study

An IoT sentinel comprises million records approximately which are labeled as positive (prone to intrusion) or negative (no intrusion). The experimental study is carried out on a dataset named IoT Sentinel. For the experimental study, amid million records, 190000(negative: 90000, and positive: 100000) records are taken into consideration. The learning phase of the proposed model DDMFO is trained with the75% of the positive and negative label records among the given input dataset were used. The rest, 25% of positive and negative records, were used to assess the performance of the classification process in regard to accuracy and false alarming. The inputs given and outcomes obtained for the different statistical metrics often used in classifier performance assessment are listed in table1.

Table 1: The inputs and outcomes of the label prediction phase

Positives (training)	75000
Negatives (training)	67500
Positives (testing)	25000
Negatives (testing)	22500
True Positives	22558
True Negatives	20070
False Positives	2430
False Negative	2442
precision	0.903
Negative Predictive Value	0.892
Accuracy	0.897
Sensitivity	0.9023
Specificity	0.892
False Positive Rate (Fall-out)	0.0977

The suggested model utilizes 47500 (benevolent: 22500, and malevolent: 25000) records in the prediction stage. The accurately predicted records were 44541 from the outcomes of predictive analysis. Of these 19853 are accurately labeled as negative. And 24688 are accurately labeled as positive. Therefore sensitivity which is defined as "true positive rate," is 0.9023(It is the ratio of true positives (TP) contrary to actual positives), and the specificity, which is defined as "true negative rate," is 0.892 (It is the ratio of true negatives (TN) contrary to actual negatives). And 2959 records is the amount of falsely predicted records. Amid these, the total number of falsely labeled records is 2647, which are considered malevolent (positive), and the total number of 312 falsely labeled records is considered as benevolent (negative). The "positive predictive value" is 0.903 (It is the ratio of TP against the aggregate of TP and FP), and the "negative predictive value" is 0.892 (It is the ratio of TN against the aggregate of

TN and FN). And 0.897 is the complete predictive accuracy which is the ratio of an aggregate of TP and TN against to the number of records utilized in predictive analysis). The analysis reveals that the suggested heuristics to measure the malevolent and benevolent extent of IoT network dealings are significant in distinguishing IoT network traffic defined as malevolent and benevolent with an accuracy of 89.7%. The represented sensitivity of the suggestion (sensitivity: 90%) signifies that the miss rate is low. Since the specificity is approx. 89%, fall out is also considerably high, which is 9%. In Figure 1, the metric values are shown:

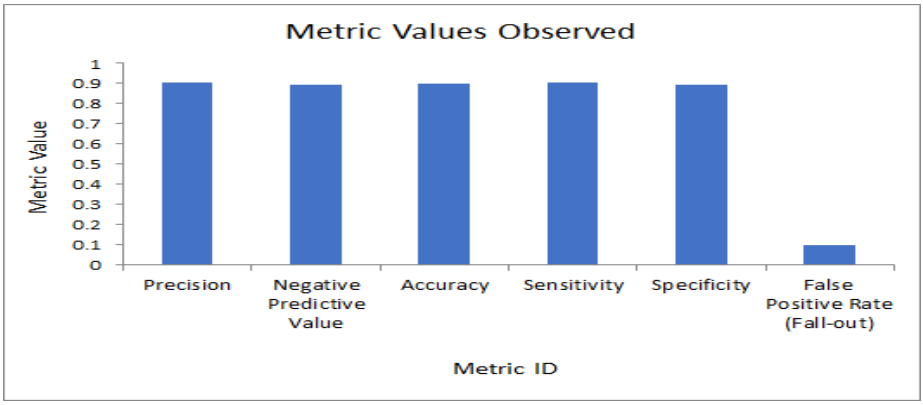

Figure 1: The metric values observed from the experimental study

5. Conclusion

The proposed system is a feature selection and optimization technique built over the statistical method called Dice similarity Coefficient. The proposal is intended to select optimal features from the set of IoT network transactions labeled as positive or negative to prone the intrusion practices. The contribution is considered significant and robust to selecting and optimizing the features in regard to training the classifier that depicts the given IoT network transaction that is vulnerable to intrusion or not. The output of the proposed model was scaled through the classifier called naïve Bayes. The Experimental investigation points out that the suggested method is highly significant. However, the false-positive rate of benevolent record forecast is generally very high. On the other hand, Extreme protection against intrusion inside sensitive IoT networks can be tolerated.

References

[1] Kun, Z.,Meng, X.: Research and prevention measures of computer network security. Fujian 10, (2009); 102–103.
[2] United States General Accounting Office: Computer Attacks at Department of Defense Pose Increasing Risks. GAO/AIMD-96-84 Defense Information Security, Washington DC (1996).
[3] United States General Accounting Office: Opportunities for improved OMB oversight of agency practices. GAO/AIMD Information Security, Washington DC (1996).
[4] Conti, J.P.: The Internet of things. Commun. Eng. 4(6), (2006); 20–25.

[5] Zhou, J.: Wireless sensor network intrusion detection model research. In: CIE 16th Information Theory Academic Conference Proceedings, Electronic Industry Press, Beijing pp. 799–804; (2009).
[6] Selvakumar, K., L. Sairamesh, and A. Kannan. "Wise intrusion detection system using fuzzy rough set-based feature extraction and classification algorithms." *International Journal of Operational Research* 35, no. 1 (2019): 87-107..
[7] Ariu, Davide, Roberto Tronci, and Giorgio Giacinto. "HMMPayl: An intrusion detection system based on Hidden Markov Models." computers & security 30.4 (2011): 221-241.
[8] Koc, Levent, Thomas A. Mazzuchi, and ShahramSarkani. "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier." Expert Systems with Applications 39.18 (2012): 13492-13500.
[9] Lin, Wei-Chao, Shih-Wen Ke, and Chih-Fong Tsai. "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors." Knowledge-based systems 78 (2015): 13-21.
[10] Thorsten, Marina, and Chuanyi Ji. "Anomaly detection in IP networks." IEEE Transactions on signal processing 51.8 (2003): 2191-2204.
[11] Weller-Fahy, David J., Brett J. Borghetti, and Angela A. Sodemann. "A survey of distance and similarity measures used within network intrusion anomaly detection." IEEE Communications Surveys & Tutorials 17.1 (2015): 70-91.
[12] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. "Network anomaly detection: methods, systems and tools." Ieee communications surveys & tutorials 16.1 (2014): 303-336.
[13] Theiler, James P., and D. Michael Cai. "Resampling approach for anomaly detection in multispectral images." Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX. Vol. 5093. International Society for Optics and Photonics, 2003.
[14] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.
[15] Casas, Pedro, Johan Mazel, and Philippe Owezarski. "Unsupervised network intrusion detection systems: Detecting the unknown without knowledge." Computer Communications 35.7 (2012): 772-783.
[16] De la Hoz, Eduardo, et al. "PCA filtering and probabilistic SOM for network intrusion detection." Neurocomputing 164 (2015): 71-81.
[17] Bostani, Hamid, and Mansour Sheikhan. "Modification of supervised OPF-based intrusion detection systems using unsupervised learning and social network concept." Pattern Recognition 62 (2017): 56-72.
[18] Guo, Chun, et al. "A two-level hybrid approach for intrusion detection." Neurocomputing 214 (2016): 391-400.
[19] Horng, Shi-Jinn, et al. "A novel intrusion detection system based on hierarchical clustering and support vector machines." Expert systems with Applications 38.1 (2011): 306-313.
[20] Tan, Zhiyuan, et al. "A system for denial-of-service attack detection based on multivariate correlation analysis." IEEE transactions on parallel and distributed systems 25.2 (2014): 447-456.
[21] Pajouh, Hamed Haddad, GholamHosseinDastghaibyfard, and SattarHashemi. "Two-tier network anomaly detection model: a machine learning approach." Journal of Intelligent Information Systems 48.1 (2017): 61-74.
[22] Osanaiye, Opeyemi, Kim-Kwang Raymond Choo, and MqheleDlodlo. "Distributed denial of service (DDoS) resilience in cloud: review and conceptual cloud DDoS mitigation framework." Journal of Network and Computer Applications 67 (2016): 147-165.
[23] Kamalanathan, Selvakumar, Sai Ramesh Lakshmanan, and Kannan Arputharaj. "Fuzzy-clustering-based intelligent and secured energy-aware routing." In Handbook of Research on Fuzzy and Rough Set Theory in Organizational Decision Making, pp. 24-37. IGI Global, 2017.
[24] Ambusaidi, Mohammed A., et al. "Building an intrusion detection system using a filter-based feature selection algorithm." IEEE transactions on computers 65.10 (2016): 2986-2998.
[25] Daryabar, Farid, et al. "Towards secure model for SCADA systems." Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2012 International Conference in. IEEE, 2012.
[26] Pan, Shengyi, Thomas Morris, and UttamAdhikari. "Developing a hybrid intrusion detection system using data mining for power systems." IEEE Transactions on Smart Grid 6.6 (2015): 3104-3113.
[27] Zhou, Chunjie, et al. "Design and analysis of multimodel-based anomaly intrusion detection systems in industrial process automation." IEEE Transactions on Systems, Man, and Cybernetics: Systems 45.10 (2015): 1345-1360.
[28] Jaithunbi, A. K., S. Sabena, and L. SaiRamesh. "Trust evaluation of public cloud service providers using genetic algorithm with intelligent rules." Wireless Personal Communications 121, no. 4 (2021): 3281-3295.
[29] Ashraf, Qazi Mamoon, and Mohamed HadiHabaebi. "Autonomic schemes for threat mitigation in the Internet of Things." Journal of Network and Computer Applications 49 (2015): 112-127.

[30] Ning, Huansheng, Hong Liu, and Laurence Yang. "Aggregated-proof based hierarchical authentication scheme for the internet of things." IEEE Transactions on Parallel & Distributed Systems 1 (2015): 1-1.
[31] Cao, Xianghui, et al. "Ghost-in-ZigBee: Energy depletion attack on ZigBee-Based wireless networks." IEEE Internet of Things Journal 3.5 (2016): 816-829.
[32] Chen, Qian, SherifAbdelwahed, and AbdelkarimErradi. "A model-based validated autonomic approach to self-protect computing systems." IEEE Internet of Things Journal 1.5 (2014): 446-460.
[33] Teixeira, F. Augusto, et al. "Defending Internet of Things against Exploits." IEEE Latin America Transactions 13.4 (2015): 1112-1119.
[34] Cunningham, Padraig. "A taxonomy of similarity mechanisms for case-based reasoning." IEEE Transactions on Knowledge and Data Engineering 21.11 (2009): 1532-1543.