Received: March 25, 2017

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Improvised *k*-NN Respecting Diversity of Data for Network Intrusion Detection

Yasir Hamid[1*], Balasaraswathi Ranganathan[1], Ludovic Journaux[2], Qaiser Farooq[3], Sugumaran Muthukumarasamy[1]

[1]*Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India*
[2] *LE2I UMR6306, CNRS, Univ.~Bourgogne Franche-Comtê, AgroSup Dijon, France*
[3]*Department of Statistics, Pondicherry Central University, Pondicherry, India*
* Corresponding author's Email: bhatyasirhamid@pec.edu

Abstract: Network Intrusion Detection is a complex classification problem aimed at discriminating the legitimate from illegitimate and potentially harmful network connections over the communication network. What adds to the complexity of the problem is the near real-time response to a threat, imbalanced datasets to deal with and finally the data being mixed in nature with some features being numeric some discrete and some nominal. In this work we have applied Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and eliminate the skewness of the class distribution. The success of *k*-Nearest Neighbour (*k*-NN) depends upon the set of neighbours deemed to be very close or similar to a data point which is in turn determined by the similarity/distance metric employed, where most of the metrics employed in literature deal with numeric data only, and either need conversion of categorical features to numeric features or simply eliminated the categorical features, which often leads to reduction in the results. As for this work is considered, we take into consideration both the categories of features simultaneously by replacing the conventional Euclidean metric with Gower metric, which is better suited for mixed data. Gower metric provides a mechanism to deal with heterogeneous features differently and ultimately yields a quantifiable value that determines the similarity of the two instances. Experimental results show that improvised version of *k*-NN outperforms its conventional counterpart in terms of the Accuracy, Detection Rate, Precision, Recall, f-Measure, and Receiver Operating Characteristic (ROC) curve.

Keywords: Gower metric, Intrusion detection, KDD'99, *k*-NN.

1. Introduction

The increased use of online services has made the security of networks/systems more important than ever before. The three main principles that ensure only the authorized and authentic personals have access to information i.e., Confidentiality, Availability, and Integrity must be upheld at all the times of system operations [1]. Due to the fact that more and more services are being put online, the Internet has become an engine of communication and on a constant basis, attackers endeavour to penetrate them to steal information [2]. Internet being laid down as a distributed commodity lacks the focal security system, and it is the responsibility of the network administrators to safeguard the interests of their organizations [3]. This quest of securing the networks from the users with destructive mind set has resulted in a lot of devices being surfaced up. Most popular of them all is automatic Intrusion Detection System (IDS). An IDS is a device against whom the responsibility of discriminating the normal and hazardous traffic traversing the network is laid on [4].

Automatic IDS has enjoyed lots of attention and acceptance due to the fact that it doesn't need human intervention which proves to be inefficient more often than not. An IDS can be categorized as Host-Based IDS (HIDS) or Network-Based IDS (NIDS) [5] based on the scope of surveillance, where HIDS is installed on individual computer systems and hence are very close to the target of the attacker. These IDS

are very effective for point attacks where the attacker aims at attacking an individual machine, but are pretty ineffective for distributed attacks. Conversely, NIDS is installed at the entry point of the network i.e., just behind the firewall and there is usually only one instance of NIDS per network setup. An ideal security mechanism should have a combination of HIDS and NIDS working in collaboration. Based on the detection methodology at the heart, an IDS can be categorized as Misuse based or Anomaly based [6], where misuse based maintains a signature base of attacks and compares the captured traffic for any match from the signature base, while as anomaly based IDS learns a model for normal data, and checks how closely the captured data resembles the learnt model [7]. If it differs by more than some threshold, then is an attack else a normal connection. Misuse based systems are very effective for detecting the known attacks but are rendered helpless while encountering new attacks or the even variations of the known attacks for that case. Anomaly based methods on the other hand are able to find out the unknown attacks but suffer from the False Positives [8, 9].

A lot of Machine Learning (ML) techniques have been applied for NID for a long time now all focusing on different aspects of IDS and improving different parameters [10, 11]. In this work we applied a simple most lazy learner i.e., *k*-NN to solve the problem at hand. *k*-NN being simple most ML algorithm doesn't have any training phase rather it delays all the processing for the testing phase. Since, the only benchmark dataset for IDS i.e., the KDD99 dataset is mixed nature with some features being numeric and some being nominal. A classification algorithm that works only in numeric or only in nominal space is bound to produce inaccurate results. We in this work tried to care for this problem by the application of Gower index, that is better-suited similarity measure for mixed data, which doesn't need the conversion of nominal to numeric features or vice versa. In this work, we applied the standard Gower where the weight for each feature is set to 1 and hence each feature is deemed equally important for classification. The setup is tested over varying neighbourhood sizes from 5 to 10 at on each neighbourhood size the improvised *k*-NN outperforms its conventional counterpart based on the Euclidean distance that is well suited for numeric data only.

The rest of the paper is organized as follows. A brief literature review about various works is given in section 2, materials and methods employed in this work are discussed in section 3. Section 4, presents the methodology of the paper. Results and discussion are given in section 5. Finally, the paper concludes in Section 6.

2. Literature review

Due to the simplicity of lazy classifiers they have been pretty popular for the classification problem, they provide an idea about how much the proposed model has gained in terms of accuracy. It is common in ML fraternity that the proposed complex model is usually tested with *k*-NN to show how much improvement was acquired as in [12, 13] where the idea of *k*-NN is just to provide an idea about he attained improvement of results. Mostly *k*-NN is complimented with some weighting function so as to avoid useless features dominate the final classification as in [14 - 16], since there is still the need of conversion of the features this negatively impacts the classification process. Mostly *k*-NN using Euclidean distance has been applied for classification which needs the nominal features to be converted in numeric features, which in turn has a negative effect on the classification. The success of *k*-NN is highly dependent on the quality of the neighbours selected which in turn is highly dependent on the similarity measure applied. Lots of similarity distance measures have been applied off late like Minkowskian [17, 18], Mahalanobis [19, 20] and Point Based [21]. All the works mentioned above make use of metrics that are effective for numeric data and often lead to lead poor performances of the heterogeneous datasets Heterogeneity is at the core of NID, since the standard dataset for Network Intrusion Detection is mixed in nature with few attributes being nominal and few numeric, this warrants of some metric that is able to respect the diversity of the dataset and is able to find out the neighbors which are indeed much similar to the starting point and hence highly effects the classification.

3. Material and methods

In the next few subsections we present an exhaustive discussion of various materials and methods applied in this work.

3.1 *k*-NN classifier

k-NN [22] is highly adaptable to versatile environment and have potential to habituate almost everything ranging from vision to bio-informatics to graphical structure and so on. It's used both in Classification and Regression problems. *k*-NN belongs to non-parametrized group of probabilistic distributions. There requires no learning procedure beforehand only when the prediction is demanded the learning begins, so its aptly called as lazy learning algorithm. It's also referred as instance based

learning due to property of training the model over raw instances.

The theory of *k*-NN is to sort out a set of training samples which are neighbouring in distance to the point to be predicted and label it accordingly. The count of samples to be considered is either user defined or will vary according to density of points. Most commonly used distance measure along *k*-NN is Euclidean distance. But other distance measures such as Hamming distance, Manhattan distance, Minkowski distance, Tanimoto distance, Jaccard distance, Mahalanobis distance and Cosine distance.

The selection of *k* value plays the crucial role in *k*-NN algorithm. In case of Regression problem, the prognosis is done by,

$$\upsilon = \frac{1}{K}\sum_{n=1}^{K} \upsilon_n \qquad (1)$$

Where υ_n is the nth case of the sample and υ is the result of the query. In case of Classification problem, the prognosis in done on voting scheme, where the label which wins the majority of votes claims the point to be under its classification.

Algorithm 1 *k* – Nearest Neighbour

procedure *k*-NN ALGORITHM
Collect the Sample data set S={a^n, b^n} *a* is the data point; *b* is the associated labelling
Calculate the K-factor value ℵ
Obtain the variable new_a which requires to be labelled
Retrieve the closest data-points of the new_a based on the K-factor ℵ and distance measure ℸ
for all *i* in *Countof* ℵ **do**:
    if X have lesser ℑ **then**:
        ℵ []=a_i
    end if
end for
Each ℑ is of threshold distance measure
Acquire the number of votes

$$\eta_i\, \Im_i = \begin{cases} \nu^1, & Class\ A \\ \nu^2, & Class\ B \\ \quad \ldots\ldots\ldots & \\ \nu^n, & Class\ N \end{cases}$$

Find the class containing the majority number of votes in eta_i which wins the label *b* to the new_a
end procedure

3.2 KDD99 dataset

The first version of KDD99 [23] dataset namely DARPA98 was generated by a group of Lincoln Laboratories at MIT University. They performed a simulation of normal and attack connections over the military network and the data traversing over the communication lines were captured. The dataset is in total comprised of 9 weeks of raw TCP dump files and is divided into two subgroups part of spared for testing and rest for training. The training data consists of 7 weeks captured being processed into approx. 5 million connections and is about 4 GB in size. The rest two weeks' data was processed into 2 million connections and used for testing purpose. Totally there are connections in the dataset pertaining to 23 different attack groups coarsely categorized in 4 broad categories i.e., Denial of Service (DOS), Probe, Remote to Local (R2L) and User to Root (U2R) attacks. The test set contains 15 different attacks from 4 different groups. Both training, as well as a test set, are drawn from different frequency distributions. For each connection vector, 41 attributes and a class label were extracted using a tool known as BRO-IDS. These 41 features are categorized into three groups i.e., intrinsic features, content features, and traffic features. The features are mixed some being numeric like some being nominal and some being binary like. Appropriate pre-processing is used to before passing this data to a classifier. Despite suffering from severe criticism for being redundant, outdated and skewed, it is still continuing to be popular for IDS classifier evaluation. Due to the finiteness of the processing capabilities of any machine, full data has been seldom used. Usually, a carefully drawn subset of the full dataset is considered, and also a lot of research has been done in selecting highly discriminative features from a set of 41 features.

3.3 SMOTE

SMOTE [24] is a synthetic oversampling technique that is aimed at producing the instances of the classes synthetically. The generated instances are created synthetically rather than simply repeating few instances in the dataset, which lead to over-fitting of the classifier as it has to deal with repetitions of few instances only. At the depth what SMOTE does it employs a *k*-NN to find out the neighbours of instances and then applies a well-defined mathematical function to generate new instances. SMOTE motivated by a procedure that demonstrated to be effective for handwritten character recognition produces engineered instances in a less application-specific manner, by working in space" rather than

"data space". For each minority class, synthetic instances are introduced along with all the lines of detected neighbourhood of size *k* which tend to be close to the minority class instances. Depending on the amount of over-sampling required, neighbours from the *k* nearest neighbours are randomly chosen. We in this work have set up the neighbourhood size to 5 and have repeatedly applied SMOTE a number of times. For instance, if the amount of over-sampling needed is 200%, only two neighbours from the five nearest neighbours are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach successfully compels the choice locale of the minority class to wind up distinctly broader. For each instance, of the original data sample [][] a set of neighbours nnarray[][] are selected a set of synthetic instances synthetic[][] is doctored along all lines. The below-given code statements depict how SMOTE actually functions.

Algorithm 2 SMOTE

1: **for** Attribute 1 to # ATTRIBUTES **do:**
2: Compute the $difference$ (Φ)between Sample[nnarray[nn]][Attribute] and Sample[i].[Attribute]
3: Compute the $Random\ number$ (φ), ranging between 0 and 1
4: Compute $Synthetic[index][Attribute](\Psi)$ which is difference between Sample[i][Attribute] and Random_Number
$*difference(\Phi)$
5: **end for**

where *index* keeps track of the number of synthetic symbols generated.

3.4 Gower metric

Off late Euclidean metric has been the most popular distance metric among the ML fraternity giving the simplicity of its calculation in addition to the solid mathematical foundation. Even though it has enjoyed a lot of popularity among statisticians, mathematicians and to some extent Machine Learning Experts also, but what is peculiar to machine learning is the mixed nature of the data. Dataset having nominal and numeric attributes at the same time is a routine matter in ML. Till now the approaches that have been reported mostly perform the conversion of the nominal attributes to the numeric attributes. Here in this paper, we skip the details about various conversion algorithms, although for reference the detailed survey of such methods can be found in [25]. The problem with the conversion methods is that there is not the implicit meaning of the assigned values to the attribute and no natural relation between attribute values and its representation. What is actually needed is some sort of metric that can deal with mixed data, and can effectively compute the distance (Similarity) between instances. Review of the literature has shown that many distance metrics for mixed data have been proposed in different studies from time to time, and the one that received lots of attentions over the decades is Gower Index, although incepted for the biological studies, has been applied to the diverse fields of classification for last few decades. Gower's Similarity Coefficient is one among the prominent measure which reveals the similarity or dissimilarity betwixt the neighbourhoods and has strong roots in the ecological study. It's quite popular due to its genuinely to measure the closeness among mixed data types. Based on the comparison of pairwise items, the Gower's coefficient will be able to credit differential weights over the similarity record obtained. As said earlier the scores can be enumerated over diversified types of characteristics including categorical (dichotomous, nominal, interval, or ratio scale, ...) or numeric (real or integer quantities)). In order to calculate Gower's Similarity Coefficient, let's take into consideration two instances namely α_i and α_j with N features, now the Gower's Coefficient is calculated as $\sum_{x=1}^{N} w_x S_x(\alpha_{ix}, \alpha_{jx})$ where w_x is a binary weight of x^{th} feature which is set to 1 if comparison is possible and 0 otherwise. $S_x(\alpha_i, \alpha_j)$ is the score function as is calculated as

$$S_x(\alpha_i, \alpha_j) = \begin{cases} \varrho(\alpha_{ix}, \alpha_{jx}) & if\ Qualitative \\ \dfrac{|\alpha_{ix} - \alpha_{jx}|}{r_x} & if\ Quantitative \end{cases} \quad (2)$$

Where $r_x = \max(\alpha_x) - \min(\alpha_x)$ normalizes the value to the range [0 – 1], and ϱ is Dirac's function, which is one iff α_i and α_j are from the same leagues. i.e.,

$$S_x(\alpha_i, \alpha_j) = \begin{cases} 0 \; if \; a_{ix} = a_{jx} \\ 1 \; if \; a_{ix} \neq a_{jx} \end{cases} \quad (3)$$

This Gower function is essence is the weighted average of the distances of different variables. Here in this work we have set $w_x = 1$ for all the features meaning that each feature has equal say in determining the class of the record.

4. Methodology

The experiments start with the subset selection of the data, as it would not be possible to use the whole of the dataset due to the computational constraints. Care was taken to maintain the representation from all the classes of the data. The drawn subset was subjected to random under sampling of the majority class and SMOTE based oversampling of the minority classes. As for this work is considered in order to apply Gower metric 7 symbolic features were converted into numeric form. The numeric for a feature was not assigned arbitrarily rather we made use of indicator variables in which a group of binary variables was used to represent each symbolic attributes. This resulted in the expansion of dataset to a total of 119 dimensions. After conversion of attributes features were scaled to [0-1] range so as to avoid features with small numeric values being dominated by the features with larger values. From here onwards the k-NN is used to classify the dataset using tenfold cross-validation, here we have replaced the traditional Euclidean distance with the Gower metric. Moreover, we have also used the balanced data without binarization as such and applied *k*-NN implementing Euclidean distance to be sure that the proposed model has actually improved the results for all the groups of attacks as well as normal data connections.

5. Results and discussion

The proposed system is evaluated using five different performance metrics over five different neighbourhood sizes. In the following few subsections we will first discuss about each metric and also mention their formulae and finally present the results of the proposed Gower *k*-NN and at the same time compare it with Simple *k*-NN i.e., applying distance measures suited for numeric data only such as Euclidean distance, Manhattan distance, Chebchev Distance, and finally the Camberra distance. The purpose of this comparison is to provide test out how the proposed change of distance metric effects the accuracy of any k-NN based IDS. As it has been already mentioned that an IDS is bound to produce very superior results if it can discover ideal neighbours that is in turn controlled by the distance measure employed. Once the distance measure is selected there is only on hyper parameter that needs to be set i.e., size of the neighborhood or in other words we can say value of k. Various distance measures employed for comparison in this paper are calculated as given in Table 1 below.

As can be seen from the table above all the distance measures documented unanimous way of calculating the distance between instances without taking into consideration the type of the attributes. Both numeric and categorical attributes are dealt with same treatment. This bogus assumption of homogenous treatment of all the types of attributes has effect on classification problems in general and ID in particular. The need was to augment the k-NN with a metric that is suited for mixed data hence Gower metric was employed.

5.1 Accuracy

Figure 1 presents the line diagram of Accuracy of the system over five different neighbourhood sizes.

Table 1. Various Distance Measures

SNO	Measure	Formula
1	Euclidean	$d(x,y) = \left(\sum_{i=1}^{m} \lvert x_i - y_i \rvert^2\right)^{1/2}$
2	Manhattan	$d(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$
3	Camberra	$d(x,y) = \sum_{i=1}^{m} \left\lvert \frac{x_i - y_i}{x_i + y_i} \right\rvert$
4	Chebchew	$d(x,y) = \max_{1 \le i \le m} \lvert x_i - y_i \rvert$

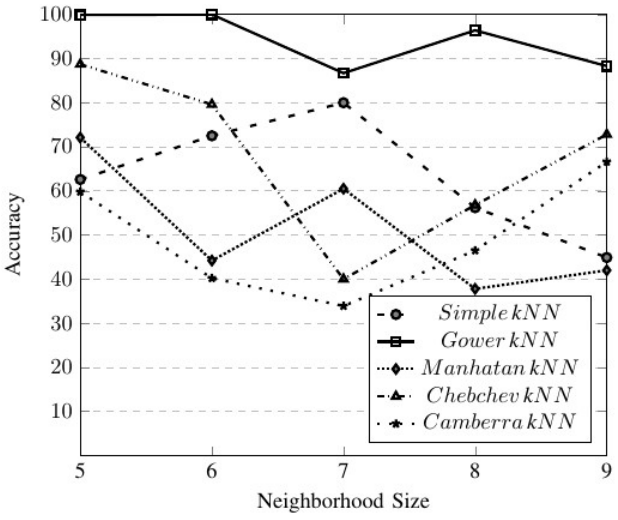

Figure.1 Accuracy

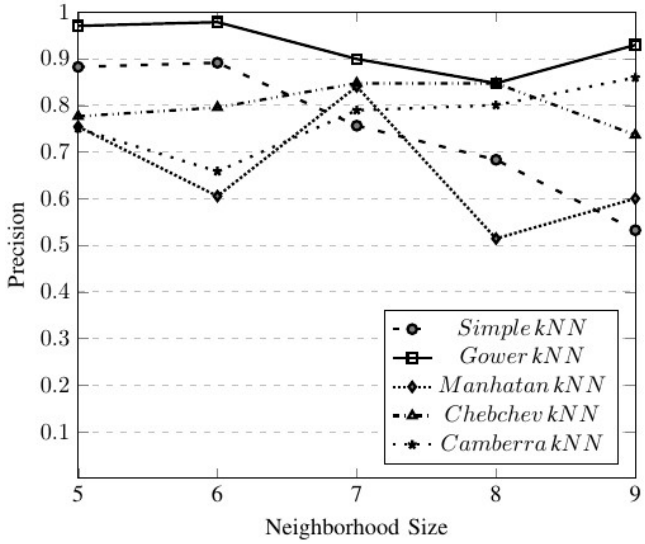

Figure. 2 Precision

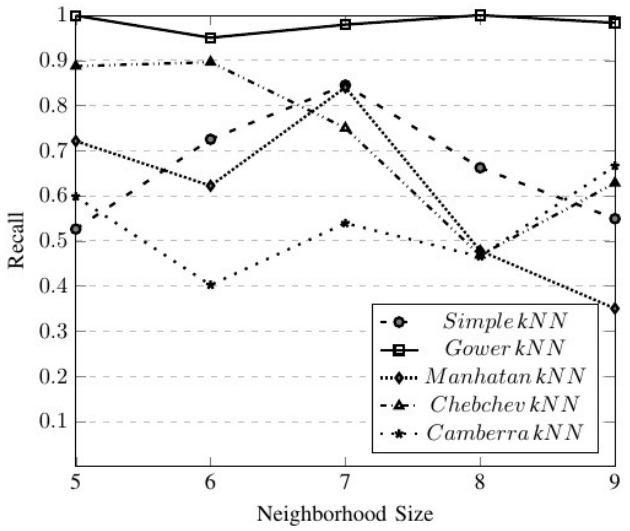

Figure. 3 Recall

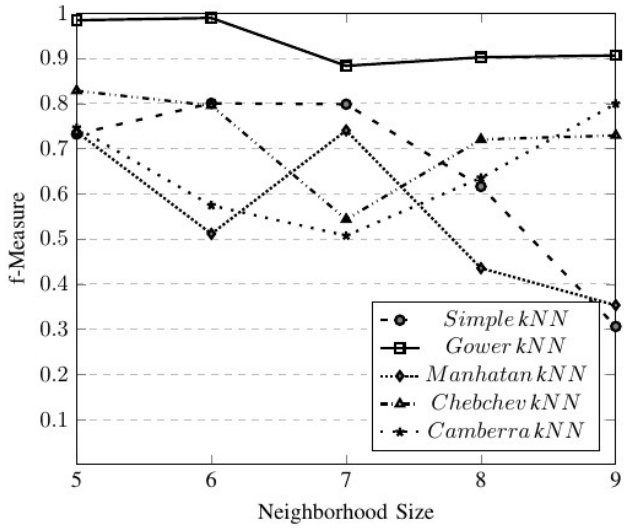

Figure. 4 f-Measure

Accuracy of a system is calculated using the following formula $Acc = (TP + TN)/(TP + FP + TN + FN)$ where TP stands for True Positive, TN stands from True Negative, FP stands for False Positive and FN stands for False Negative.

As can be seen from the figure above Gower *k*-NN yields better accuracy than normal *k*-NN over all neighbourhood sizes. Only when neighbourhood size is set to 7 both the models report same accuracy, from that point onward the accuracy of simple k-NN drops again but that of improvised k-again tends to increase, because of the fact Gower metric gives good measure of similarity.

5.2 Precision

Figure 2 presents the line diagram for precision of simple *k*-NN and Gower *k*-NN over varying neighbourhood sizes. The precision of any system is calculated as $Precision = TP/(TP + FP)$. As can be seen from the figure for all the neighbourhood sizes from 5 to 9 improvised k-NN yields higher Precision than the simple *k*-NN. At all the values of k there is a considerable difference between the Precision of the two models, not even on one k value the simple *k*-NN has as good Precision as the improvised *k*-NN.

5.3 Recall

Figure 3 presents the line diagram for recall of simple *k*-NN and Gower *k*-NN over varying neighbourhood sizes. The precision of any system is calculated as $Recall = TP/(TP + FN)$. As can be seen from the figure almost on all the values of k improvised k-NN has better recall than simple k-NN with Euclidean distance. Only at one value of k the two systems yield same Recall.

5.4 f-measure

Figure 4 presents a line diagram of f-Measure for simple k-NN and Gower k-NN over varying neighbourhood sizes. f-Measure provides simple measure that combines Precision and Recall to a single number, mathematically it is given as $2\left(\frac{Precision\,.\,Recall}{Precision+Recall}\right)$. As can be seen from the figure given below Gower *k*-NN yields very high f-Measure than simple *k*-NN over all the neighbourhood sizes. The difference between the two is considerable at all the k values.

5.5 ROC

Figure 5 presents a line diagram of Area under Curve of ROC for simple *k*-NN and Gower *k*-NN over five different neighbourhood sizes. ROC graphs

are two-dimensional graphs in which TP rate is plotted on the Y axis and FP rate is plotted on the X axis. An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). Since ROC are actually curves in the real life to use them for measuring the effectiveness of classifier we calculate area under curve. As it is pretty evident from the figure given below that on all but one k values Gower *k*-NN has better ROC value than simple *k*-NN.

As can be seen from the discussion above that the improvised k-NN with Gower metric outperforms all other versions of k-NN on almost all neighborhood sizes. The reason for this appreciable improvement is attributes to the fact that the success of k-NN is determined by the quality of its neighbourhood which is in turn determined by the distance/similarity function employed. Any metric that needs the conversion of the attributes or simply discards the heterogeneity is bound to sacrifice on the results. Hence the metrics suited for homogenous data perform pretty poorly when put to deal with heterogeneous data. From the results we can easily claim that the proposed model is capable of respecting the diversity of the data and hence enjoys the improved results in terms of Accuracy, Precision, Recall, f-Measure and ROC.

As it is pretty clear by now that there are attacks pertaining to different groups in the KDD99 dataset. Not just attacks, there are connections in the datasets that represent normal transactions. An ideal IDS is one that has acceptable detection rate for all the groups of attacks as with the least false positive rate. What that means that IDS should be capable of detecting the normal data connections effectively and should not be blocked on the suspicion of being attacks.

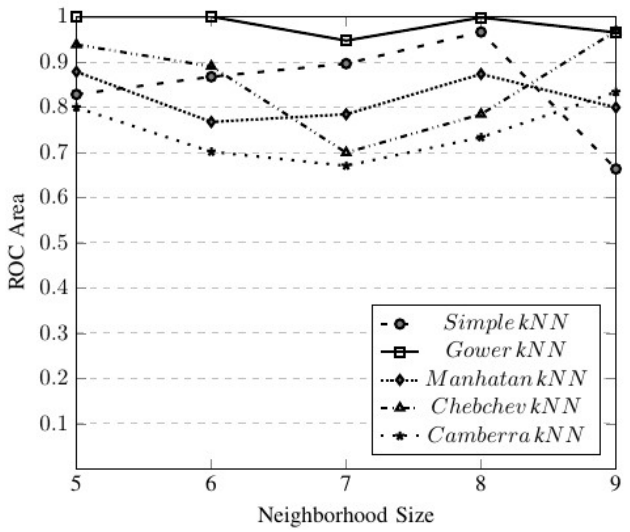

Figure. 5 ROC Area

Table 1. Comparison with Related Works

MODEL	NORMAL	DOS	PROBE	R2L	U2R
Chen et al. [26]	99.50	97.60	91.40	90.30	53.81
Wang et al. [27]	97.94	97.50	76.38	15.38	09.77
Toosi et al. [28]	98.20	99.50	84.10	31.50	14.20
CF Tsai et al. [29]	96.12	83.12	96.59	78.95	61.54
Simple k-NN	95.80	92.70	81.80	57.45	57.80
Gower k-NN	**99.96**	**99.89**	**99.60**	**95.96**	**70.64**

Since the problem at hand is a multiclass classification problem and classifier model should be able to effectively detect all the groups of attack. Actually, in total, there is a group of 23 attacks in the dataset, these attacks as already mentioned in categorized in four different broad groups i.e, DoS, Probe, R2L and U2R. There have been few works that have reported the classification for each group of the attacks of the dataset. Likewise, we also in this work have tried to check out how the proposed method work at the category level. A comparison of proposed work with some of the proposed method with some of the prominent works in the field that have mentioned put the results at each category level is given in the Table 1.

As can be seen from the table given above, no reported work has equally appreciable detection rate for all the groups of attacks. The reason for the underperformance of the models may not be pinned down to one, many of them can affect the detection rate Class Imbalance, Skewness, and unsuitable metric being a few. We in this work have taken care of most of the problem that results in the lower detection rate for minimal classes, and this effort has indeed improved the results as can be easily concluded from the table. Still, there is some scope for improvement for U2R and R2L attack group. Which we believe can be attained by using variable weight mechanism for Gower metric as the present setup assigns equal weights to all features (i.e., for All Wx=1), this bias handicaps *k*-NN allowing redundant, irrelevant and other imperfect features to influence the distance computation, the presence of such features would more often than not prove to be detrimental for classification which can be eliminated by controlling the influence of such variables of such features on the classification.

6. Conclusion

We in this paper have presented an improvised of a *k*-NN classifier that is aimed at respecting the diversity of the data for network intrusion detection. The dataset for network intrusion detection being mixed in nature caused bias in the classifier when Euclidean distance measure is used. Gower metric which is better suited for the mixed data was applied to replace the conventional and highly popular Euclidean distance. Experiments were run on varying neighbourhood sizes from five to ten, and for each neighbourhood size, the results were checked using 10-fold cross validation. Results have shown that on almost all the neighbourhood sizes our improvised *k*-NN performs better than its conventional counterpart. The reason for this appreciable improvement is attributes to the fact that the success of k-NN is determined by the quality of its neighbourhood which is in turn determined by the distance/similarity function employed. Any metric that needs the conversion of the attributes or simply discards the heterogeneity is bound to sacrifice on the results. Here in this work we have employed the Gower metric which is much suited for the mixed data as it provides an integrated mechanism to deal with the heterogeneous data simultaneously, the effects of which are very much clear. This can be seen from the result section where the proposed model work appreciably better than the existing systems suitable for homogenous systems. The proposed system can be employed in any heterogeneous classification environment with acceptable performance. As for this work is concerned we have made an assumption that all the features are equally essential for classification but that is not true in reality, in a case there might be some features that are useful but less important than others. Assigning equal weights to all the features is practically little inappropriate. As an extension of this work, we would like to see the effects of different feature weights of the Gower metric and in addition to reduced time complexity, we would try to make use of similarity tree and similarity caching methods.

References

[1] W. R. Cheswick, S. M. Bellovin, and A. D. Rubin, *Firewalls and Internet security: repelling the wily hacker,* Addison-Wesley Longman Publishing Co., Inc., 2003.

[2] W. Stallings, *Network and internetwork security: principles and practice*, Prentice Hall Englewood Cliffs, Vol.1, 1995.

[3] H.J. Liao, C.H. R. Lin, Y.C. Lin, and K.Y. Tung, "Intrusion detection system: A comprehensive review*", Journal of Network and Computer Applications*, Vol. 36, No. 1, pp. 16–24, 2013.

[4] C. Kaufman, R. Perlman, and M. Speciner, *Network security: private communication in a public world*, Prentice Hall Press, 2002.

[5] E. Biermann, E. Cloete, and L. M. Venter, "A comparison of intrusion detection systems", *Computers & Security*, Vol. 20, No. 8, pp. 676–683, 2001.

[6] D. Anderson, T. Frivold, and A. Valdes, *"Next-generation intrusion detection expert system (nides): A summary"*, 1995.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey", *ACM computing surveys (CSUR)*, Vol. 41, No. 3, p. 15, 2009.

[8] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, "An intelligent intrusion detection system (ids) for anomaly and misuse detection in computer networks", *Expert systems with Applications*, Vol. 29, No. 4, pp. 713–722, 2005.

[9] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maci´a-Fern´andez, and E. V´azquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", *Computers & Security,* Vol. 28, No. 1, pp. 18–28, 2009.

[10] Y. Hamid, M. Sugumaran, and V. Balasaraswathi, "Ids using machine learning-current state of art and future directions", *British Journal of Applied Science & Technology*, Vol. 15, No. 3, 2016.

[11] C.F. Tsai, Y.F. Hsu, C.Y. Lin, and W.Y. Lin, "Intrusion detection by machine learning: A review", *Expert Systems with Applications*, Vol. 36, No. 10, pp. 11 994–12 000, 2009.

[12] Y. Liao and V.R. Vemuri, *"Use of k-nearest neighbor classifier for intrusion detection",* Computers & security, Vol. 21, No. 5, pp. 439–448, 2002.

[13] Y. Li and L. Guo, "An active learning based tcm-knn algorithm for supervised network intrusion detection*", Computers & security,* Vol. 26, No. 7, pp. 459–467, 2007.

[14] J.M. Keller, M.R. Gray, and J.A. Givens, "A fuzzy k-nearest neighbor algorithm", *IEEE transactions on systems, man, and cybernetics*, No. 4, pp. 580–585, 1985.

[15] T. Mohri and H. Tanaka, "An optimal weighting criterion of case indexing for both numeric and symbolic attributes", In: *AAAI-94 Workshop Program: Case-Based Reasoning,* Working Notes, pp. 123–127, 1994.

[16] R. Kohavi, P. Langley, and Y. Yun, "The utility of feature weighting in nearest-neighbor algorithms", In: *Proc. of the Ninth European Conference on Machine Learning,* Citeseer, pp. 85–92, 1997.

[17] Tversky, "Features of similarity." *Psychological review*, Vol. 84, No. 4, p. 327, 1977.
[18] Y. Biberman, "A context similarity measure," In: *Proc. of European Conference on Machine Learning*, Springer, pp. 49–63, 1994.
[19] S. Zhang and X. Pan, "A novel text classification based on Mahalanobis distance," In: *Proc. of 3rd International conference on Computer Research and Development (ICCRD)*, Vol. 3, pp. 156–15, IEEE, 2011.
[20] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, Vol. 41, No. 12, pp. 3600–3612, 2008.
[21] C. G. Atkeson, *"Using local models to control movement"* In: NIPS, pp. 316–323, 1989.
[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, Vol. 13, No. 1, pp. 21–27, 1967.
[23] K. Cup, *"Available on: http://kdd. ics. uci.edu/databases/kddcup99/kddcup99.html"*, 2007.
[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, Vol. 16, pp. 321–357, 2002.
[25] J. C. Gower, "A general coefficient of similarity and some of its properties", *Biometrics*, pp. 857–871, 1971.
[26] C.M. Chen, Y.L. Chen, and H.C. Lin, "An efficient network intrusion detection", *Computer Communications*, Vol. 33, No. 4, pp. 477–484, 2010.
[27] S.S. Wang, K.Q. Yan, S.C. Wang, and C.W. Liu, "An integrated intrusion detection system for cluster-based wireless sensor networks", *Expert Systems with Applications*, Vol. 38, No. 12, pp. 15 234–15 243, 2011.
[28] N. Toosi and M. Kahani, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers", *Computer communications*, Vol. 30, No. 10, pp. 2201–2212, 2007.
[29] C.F. Tsai and C.Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection", *Pattern recognition*, Vol. 43, No. 1, pp. 222–229, 2010.