ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

# A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms

**L.Dhanabal[1], Dr. S.P. Shantharajah[2]**

Assistant Professor [SRG], Dept. of Computer Applications, Kumaraguru College of Technology, Coimbatore, India[1]

Professor, Department of MCA, Sona College of Technology, Salem, India[2]

**Abstract**: Intelligent intrusion detection systems can only be built if there is availability of an effective data set. A data set with a sizable amount of quality data which mimics the real time can only help to train and test an intrusion detection system. The NSL-KDD data set is a refined version of its predecessor KDD'99 data set. In this paper the NSL-KDD data set is analysed and used to study the effectiveness of the various classification algorithms in detecting the anomalies in the network traffic patterns. We have also analysed the relationship of the protocols available in the commonly used network protocol stack with the attacks used by intruders to generate anomalous network traffic. The analysis is done using classification algorithms available in the data mining tool WEKA. The study has exposed many facts about the bonding between the protocols and network attacks.

**Keywords**: Intrusion Detection System, NSL-KDD dataset, Anomaly, Protocol.

## I. INTRODUCTION

Communication system plays an inevitable role in common man's daily life. Computer networks are effectively used for business data processing, education and learning, collaboration, widespread data acquisition and entertainment. The computer network protocol stack that is in use today was developed with a motive to make it transparent and user friendly. This lead to the development of a robust communication protocol stack. The flexibility of the protocol has made it vulnerable to the attacks launched by the intruders. This makes the requirement for the computer networks to be continuously monitored and protected. The monitoring process is automated by an intrusion detection system (IDS) [1]. The IDS can be made of combination of hardware and software.

At any point of time a web server can be visited by many clients and they naturally produce heavy traffic. Each network connection can be visualized as a set of attributes. The traffic data can be logged and be used to study and classify in to normal and abnormal traffic. In order to process the voluminous database, machine learning techniques can be used.

Data mining is the process of extracting interested data from voluminous data sets using machine learning techniques [2].

In this paper the analysis of the NSL-KDD data set [3] is made by using various clustering algorithms available in the WEKA [4] data mining tool. The NSL-KDD data set is analyzed and categorized into four different clusters depicting the four common different types of attacks. An in depth analytical study is made on the test and training data set. Execution speed of the various clustering algorithms is analysed. Here the 20% train and test data set is used. This paper uses the NSL-KDD data set to reveal the most vulnerable protocol that is frequently used intruders to launch network based intrusions.

The rest of the paper is organized as follows: Section II presents some related work based on intrusion detection. Section III gives a brief description on the contents of the NSL-KDD dataset. Section IV summarizes about analysis of the dataset with various classification techniques. Section V presents the graphical analysis report on various intrusions using different classification methods. Section VI, deals with conclusion and future work.

## II. RELATED WORK

The NSL-KDD data set is the refined version of the KDD cup99 data set [5]. Many types of analysis have been carried out by many researchers on the NSL-KDD dataset employing different techniques and tools with a universal objective to develop an effective intrusion detection system. A detailed analysis on the NSL-KDD data set using various machine learning techniques is done in [6] available in the WEKA tool. K-means clustering algorithm uses the NSL-KDD data set [7] to train and test various existing and new attacks. A comparative study on the NSL-KDD data set with its predecessor KDD99 cup data set is made in [8] by employing the Self Organization Map (SOM) Artificial Neural Network. An exhaustive analysis on various data sets like KDD99, GureKDD and NSL-KDD are made in using various data mining based machine learning algorithms like Support Vector Machine (SVM), Decision Tree, K-nearest neighbor, K-Means and Fuzzy C-Mean clustering algorithms.

## III.DATASET DESCRIPTION

The inherent drawbacks in the KDD cup 99 dataset [9] has been revealed by various statistical analyses has affected

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

the detection accuracy of many IDS modelled by researchers. NSL-KDD data set [3] is a refined version of its predecessor.
It contains essential records of the complete KDD data set. There are a collection of downloadable files at the disposal for the researchers. They are listed in the Table 1

TABLE I : LIST OF NSL-KDD DATASET FILES AND THEIR DESCRIPTION

| S.No. | Name of the file | Description |
|---|---|---|
| 1 | KDDTrain+.ARFF | The full NSL-KDD train set with binary labels in ARFF format |
| 2 | KDDTrain+.TXT | The full NSL-KDD train set including attack-type labels and difficulty level in CSV format |
| 3 | KDDTrain+_20Percent.ARFF | A 20% subset of the KDDTrain+.arff file |
| 4 | KDDTrain+_20Percent.TXT | A 20% subset of the KDDTrain+.txt file |
| 5 | KDDTest+.ARFF | The full NSL-KDD test set with binary labels in ARFF format |
| 6 | KDDTest+.TXT | The full NSL-KDD test set including attack-type labels and difficulty level in CSV format |
| 7 | KDDTest-21.ARFF | A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21 |
| 8 | KDDTest-21.TXT | A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21 |

1. Redundant records are removed to enable the classifiers to produce an un-biased result.
2. Sufficient number of records is available in the train and test data sets, which is reasonably rational and enables to execute experiments on the complete set.
3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

In each record there are 41 attributes unfolding different features of the flow and a label assigned to each either as an attack type or as normal.
The details of the attributes namely the attribute name, their description and sample data are listed in the Tables II, III, IV, V. The Table VI contains type information of all the 41 attributes available in the NSL-KDD data set.
The 42nd attribute contains data about the various 5 classes of network connection vectors and they are categorized as one normal class and four attack class. The 4 attack classes are further grouped as DoS, Probe, R2L and U2R. The description of the attack classes.

TABLE II: BASIC FEATURES OF EACH NETWORK CONNECTION VECTOR

| Attribute No. | Attribute Name | Description | Sample Data |
|---|---|---|---|
| 1 | Duration | Length of time duration of the connection | 0 |
| 2 | Protocol_type | Protocol used in the connection | Tcp |
| 3 | Service | Destination network service used | ftp_data |
| 4 | Flag | Status of the connection – Normal or Error | SF |
| 5 | Src_bytes | Number of data bytes transferred from source to destination in single connection | 491 |
| 6 | Dst_bytes | Number of data bytes transferred from destination to source in single connection | 0 |
| 7 | Land | if source and destination IP addresses and port numbers are equal then, this variable takes value 1 else 0 | 0 |
| 8 | Wrong_fragment | Total number of wrong fragments in this connection | 0 |
| 9 | Urgent | Number of urgent packets in this connection. Urgent packets are packets with the urgent bit activated | 0 |

TABLE III : CONTENT RELATED FEATURES OF EACH NETWORK CONNECTION VECTOR

| Attribute No. | Attribute Name | Description | Sample Data |
|---|---|---|---|
| 10 | Hot | Number of 'hot' indicators in the content such as: entering a system | 0 |


DOI 10.17148/IJARCCE.2015.4696

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

| | | | |
|---|---|---|---|
| | | directory, creating programs and executing programs | |
| 11 | Num_failed_logins | Count of failed login attempts | 0 |
| 12 | Logged_in | Login Status : 1 if successfully logged in; 0 otherwise | 0 |
| 13 | Num_compromised | Number of ``compromised'' conditions | 0 |
| 14 | Root_shell | 1 if root shell is obtained; 0 otherwise | 0 |
| 15 | Su_attempted | 1 if ``su root'' command attempted or used; 0 otherwise | 0 |
| 16 | Num_root | Number of ``root'' accesses or number of operations performed as a root in the connection | 0 |
| 17 | Num_file_creations | Number of file creation operations in the connection | 0 |
| 18 | Num_shells | Number of shell prompts | 0 |
| 19 | Num_access_files | Number of operations on access control files | 0 |
| 20 | Num_outbound_cmds | Number of outbound commands in an ftp session | 0 |
| 21 | Is_hot_login | 1 if the login belongs to the ``hot'' list i.e., root or admin; else 0 | 0 |
| 22 | Is_guest_login | 1 if the login is a ``guest'' login; 0 otherwise | 0 |

TABLE IV : TIME RELATED TRAFFIC FEATURES OF EACH NETWORK CONNECTION VECTOR

| Attribute No. | Attribute Name | Description | Sample Data |
|---|---|---|---|
| 23 | Count | Number of connections to the same destination host as the current connection in the past two seconds | 2 |
| 24 | Srv_count | Number of connections to the same service (port number) as the current connection in the past two seconds | 2 |
| 25 | Serror_rate | The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count (23) | 0 |
| 26 | Srv_serror_rate | The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24) | 0 |
| 27 | Rerror_rate | The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23) | 0 |
| 28 | Srv_rerror_rate | The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24) | 0 |
| 29 | Same_srv_rate | The percentage of connections that were to the same service, among the connections aggregated in count (23) | 1 |
| 30 | Diff_srv_rate | The percentage of connections that were to different services, among the connections aggregated in count (23) | 0 |

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

| | | | |
|---|---|---|---|
| 31 | Srv_diff_host_rate | The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24) | 0 |

TABLE V: HOST BASED TRAFFIC FEATURES IN A NETWORK CONNECTION VECTOR

| Attribute No. | Attribute Name | Description | Sample Data |
|---|---|---|---|
| 32 | Dst_host_count | Number of connections having the same destination host IP address | 150 |
| 33 | Dst_host_srv_count | Number of connections having the same port number | 25 |
| 34 | Dst_host_same_srv_rate | The percentage of connections that were to the same service, among the connections aggregated in dst_host_count (32) | 0.17 |
| 35 | Dst_host_diff_srv_rate | The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32) | 0.03 |
| 36 | Dst_host_same_src_port_rate | The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33) | 0.17 |
| 37 | Dst_host_srv_diff_host_rate | The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33) | 0 |
| 38 | Dst_host_serror_rate | The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32) | 0 |
| 39 | Dst_host_srv_serror_rate | The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33) | 0 |
| 40 | Dst_host_rerror_rate | The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32) | 0.05 |
| 41 | Dst_host_srv_rerror_rate | The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33) | 0 |

The attack classes present in the NSL-KDD data set are grouped into four categories [5][9] :

1. DOS: Denial of service is an attack category, which depletes the victim's resources thereby making it unable to handle legitimate requests – e.g. syn flooding. Relevant features: "source bytes" and "percentage of packets with errors"
2. Probing: Surveillance and other probing attack's objective is to gain information about the remote victim e.g. port scanning. Relevant features: "duration of connection" and "source bytes"
3. U2R: unauthorized access to local super user (root) privileges is an attack type, by which an attacker uses a normal account to login into a victim system and tries to gain root/administrator privileges by exploiting some vulnerability in the victim e.g. buffer overflow attacks. Relevant features: "number of file creations" and "number of shell prompts invoked,"

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

4. R2L: unauthorized access from a remote machine, the attacker intrudes into a remote machine and gains local access of the victim machine. E.g. password guessing Relevant features: Network level features – "duration of connection" and "service requested" and host level features - "number of failed login attempts"

TABLE VI: ATTRIBUTE VALUE TYPE

| Type | Features |
|---|---|
| Nominal | Protocol_type(2), Service(3), Flag(4) |
| Binary | Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21),, is_guest_login(22) |
| umeric | Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29) diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41) |

The specific types of attacks are classified into four major categories. The table VII shows this detail.

TABLE VII : MAPPING OF ATTACK CLASS WITH ATTACK TYPE

| Attack Class | Attack Type |
|---|---|
| DoS | Back, Land, Neptune, Pod, Smurf,Teardrop,Apache2, Udpstorm, Processtable, Worm (10) |
| Probe | Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6) |
| R2L | Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16) |
| U2R | Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7) |

The Table VIII shows the distribution of the normal and attack records available in the various NSL-KDD datasets.

TABLE VIII: DETAILS OF NORMAL AND ATTACK DATA IN DIFFERENT TYPES OF NSL-KDD DATA SET

| Data Set Type | Total No. of | | | | | |
|---|---|---|---|---|---|---|
| | Records | Normal Class | DoS Class | Probe Class | U2R Class | R2L Class |
| KDD Train+ 20% | 25192 | 13449 | 9234 | 2289 | 11 | 209 |
| | | 53.39 % | 36.65 % | 9.09% | 0.04 % | 0.83 % |
| KDD Train+ | 125973 | 67343 | 45927 | 11656 | 52 | 995 |
| | | 53.46 % | 36.46 | 9.25% | 0.04 % | 0.79 % |
| KDD Test+ | 22544 | 9711 | 7458 | 2421 | 200 | 2754 |
| | | 43.08 % | 33.08 % | 10.74 % | 0.89 % | 12.22 % |

Figure 1 clearly exhibits the count of normal and various attack class records in the different train and test NSL-KDD data sets.

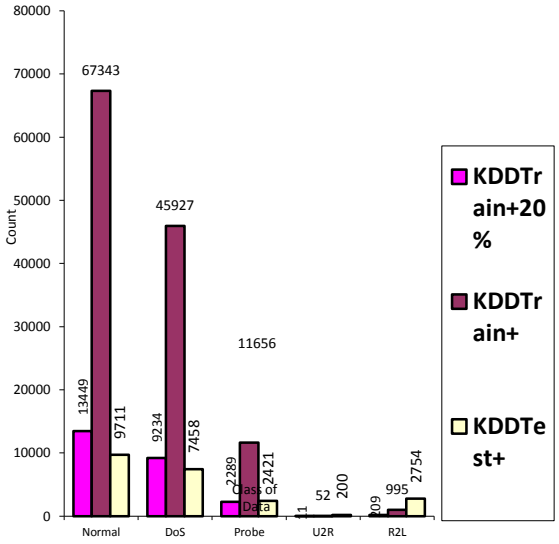


Fig 1. Network vector distribution in various NSL-KDD train and test data set

Further analysis of the KDDTrain+ data set has exposed one of the very important facts about the attack class network vectors as shown in Table IX.

From the Figure2, it is apparent that most of the attacks launched by the attackers use the TCP protocol suite. The transparency and ease of use of the TCP protocol is exploited by attackers to launch network based attacks on the victim computers.


DOI 10.17148/IJARCCE.2015.4696

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

TABLE IX: PROTOCOLS USED BY VARIOUS ATTACKS

| Attack Class / Protocol | DoS | Probe | R2L | U2R |
|---|---|---|---|---|
| TCP | 42188 | 5857 | 995 | 49 |
| UDP | 892 | 1664 | 0 | 3 |
| ICMP | 2847 | 4135 | 0 | 0 |

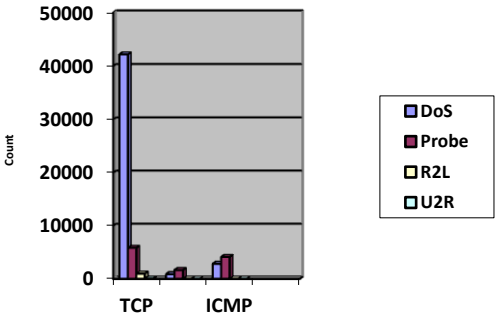


Fig2 . Protocol-wise attacks in the KDDTrain+ data set

## IV. CLASSIFICATION TECHNIQUES

Classification is a data mining method of assigning data instances in to one among the few categories. There are many classification algorithms developed to outperform one another. They all work based on mathematical techniques like decision tree, linear programming and neural networks. These techniques analyze the available data in a several ways to make its prediction.

Decision Tree: This technique divides the classification problem in to sub-problems. It builds a decision tree which in turn is used to develop a model that is used for the classification purpose.

Neural Networks: It is a collection of statistical learning models motivated by biological neural networks which are used to estimate or approximate functions that usually depend on a large amount of training data

Nearest Neighbour: This method saves all classes supplied to it by means of training data set and classifies new classes based on a similarity measure.

All the methods discussed are known for their salient features and inherent drawbacks. Decision tree takes time to build the tree. Nearest Neighbour method is considerably time consuming when the size of the data set grows. Neural network works best only on numerical data, which requires conversion of the textual data in the data set to a numerical value.

The drawbacks mentioned in the above methods give rise to an idea of going for a hybridized approach involving some optimization technique. Hybridization should consider only the salient features of the existing algorithm that could work well in the problem domain and with the available data set.

A good collection of classification algorithm with proven results are C4.5, K-Means Algorithm, Support Vector Machines (SVM), Apriori Algorithm, PageRank, AdaBoost, K-Nearest Neighbor and Naïve Bayes in existence[10].

## V . EXPERIMENTAL RESULT AND ANALYSIS

This section deals about the experiment setup and result analysis

### *A. Experiment Setup*

Many standard data mining process such as data cleaning and pre-processing, clustering, classification, regression, visualization and feature selection are already implemented in WEKA. The automated data mining tool WEKA is used to perform the classification experiments on the 20% NSL-KDD dataset. The data set consists of various classes of attacks namely DoS, R2L, U2R and Probe.

### *B. Pre-processing, Feature Selection and Classification*

The data set to be classified is initially pre-processed and normalized to a range 0 -1. This is done as a requirement because certain classifiers produce a better accuracy rate on normalized data set. Correlation based Feature Selection method is used in this work to reduce the dimensionality of the features available in the data set from 41 to 6. Classification is done in this work by using J48, SVM and Naïve Bayes algorithms

### *C. Result Analysis*

The experiments are carried out in WEKA and effectiveness of the classification algorithms in classifying the NSL-KDD data set is analyzed. The accuracy rate in detecting normal and attack class of network connection is shown in the table VI. This shows that when CFS is used for dimensionality reduction, J48 classifies the data set with a better accuracy rate.

TABLE X: ACCURACY IN DETECTION OF NORMAL AND ATTACK NETWORK FLOWS BY USING THE J48, SVM AND NAÏVE BAYES CLASSIFIERS

| Classification Algorithm | Class Name | Test Accuracy with 6 features |
|---|---|---|
| J48 | Normal | 99.8 |
| | DoS | 99.1 |
| | Probe | 98.9 |
| | U2R | 98.7 |
| | R2L | 97.9 |
| SVM | Normal | 98.8 |
| | DoS | 98.7 |
| | Probe | 91.4 |
| | U2R | 94.6 |
| | R2L | 92.5 |
| Naïve Bayes | Normal | 74.9 |
| | DoS | 75.2 |
| | Probe | 74.1 |
| | U2R | 72.3 |
| | R2L | 701.1 |

The protocol-wise distribution of normal and attack records in the various versions of the NSL-KDD data set is listed in the table VII

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 6, June 2015*

TABLE XI : PROTOCOL-WISE DISTRIBUTION OF DATA IN THE NSL-KDD DATA SET

| Data Set | TCP | UDP | ICMP |
|---|---|---|---|
| KDDTrain+20% | 20526 | 3011 | 1655 |
| KDDTrain+ | 102689 | 14993 | 8291 |
| KDDTest+ | 18880 | 2621 | 1043 |

## VI. CONCLUSION AND FUTURE WORK

The analysis results on the NSL-KDD dataset show that it is a best candidate data set to simulate and test the performance of IDS. The CFS method for dimensionality reduction reduces the detection time and increase the accuracy rate. This analysis conducted on the NSL-KDD dataset with the help of figures and tables helps the researcher to have clear understanding of the dataset. It also brings to light that most of the attacks are launched using the inherent drawbacks of the TCP protocol.

In future, it is proposed to conduct an exploration on the possibility of employing optimizing techniques to develop an intrusion detection model having a better accuracy rate.

## REFERENCES

[1] Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, Jaideep Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection"
[2] http://en.wikipedia.org/wiki/Data_mining
[3] http://nsl.cs.unb.ca/NSL-KDD/
[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)
[6] S. Revathi, Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 12, December - 2013
[7] Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013
[8] Santosh Kumar Sahu Sauravranjan Sarangi Sanjaya Kumar Jena, " A Detail Analysis on Intrusion Detection Datasets", 2014 IEEE International Advance Computing Conference (IACC)
[9] Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh," Detection of Attacks in an Intrusion Detection System", International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011, 982-986
[10] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems Journal,Springer-Verlag London, vol. 14, Issue 1, pp. 1-37, 2007.