International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

NETWORK INTRUSION DETECTION SYSTEM BASED ON MODIFIED RANDOM FOREST CLASSIFIERS FOR KDD CUP-99 AND NSL-KDD DATASET

Prakash Chandra[1], Prof. Umesh Kumar Lilhore[2], Prof. Nitin Agrawal[3]

M. Tech. Research Scholar[1], Associate Professor and Head PG[2], Associate Professor[3]
NRI-IIST Bhopal (M.P), India

Abstract - Due to rapid growth of cyber technologies number of internet users and e-data has been increased, which attracts attackers. To protect and prevent from attacks an intrusion detection system plays a vital role in cyber security. It helps to protect a network and their resources from various security threats. IDS systems are based various categories such as host based, network based and hybrid type. Networks based IDS monitors all the activities, events of the network and also analyze them for security. For detecting any intrusion in the network, a network based IDS classified the network traffic in to two classes one is normal and another is abnormal. Various data mining and machine learning methods are suggested by different researchers. Kdd cup-99 and NSL-Kdd data sets are widely used data set for the evaluation of these systems. Random forest is a widely used method in NIDS. Random forest is one of the most successful classifiers and ensemble learning. It is a classification algorithm obtained through extending the decision tree classifiers using ensemble learning techniques.

In this research paper we are presenting a network intrusion detection system method based on Modified Random forest classifiers. The existing original random forest (RF) algorithm has some challenges in feature selection process, selection of classifiers, choosing random features for various training and also challenges in combination steps. Proposed Modified random forest algorithm (MRFA) is combination of unpruned classifiers and CART (regression tree) with bagging approach. From the selected features, MRFA selects the best features and built the decision tree, a sampling variable and confusion matrix are used for identification of data more accurately and efficiently. Proposed MRFA method is implemented by using Java and simulated on weka tool and tested on Kdd-Cup 99 and NSL-Kdd dataset, and compared with Naive bayes, J-48 and Random forest. Various performance measuring parameters such as true positive rate, false positive rate, precision and F-measure are calculated. An experimental study clearly shows that proposed MRFA shows better results over existing methods.

Key Words: Intrusion detection system, NIDS, Weka, MRFA, Random forest, J-48, Navie-Bysein

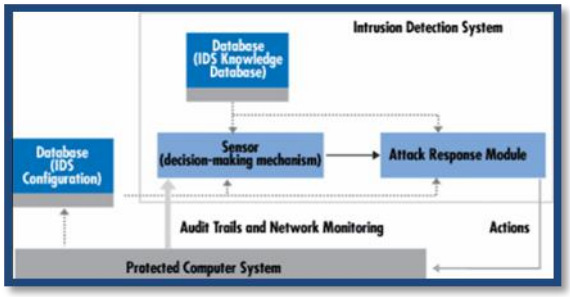
1. INTRODUCTION

In the age of computer science digital data transmission plays a vital role in communication [1]. To detect the intrusion activity, various tools like antivirus, firewall and Intrusion Detection System (IDS) are used in industry as well as in research organizations. Out of these, IDS is more powerful due to its detection capabilities within the domain of intrusion-detection there are two types of mechanism have been analyzed one is the misuse detection and second is the IDS Detection. Major concept for the misuse detection approach is to describe the attacks with the pattern or in the form of a signature such that there are even different types of attacks these approaches may get detected [11].

Intrusion detection System (IDS) is a type of security management system for computers and networks. Dependent on these types of signatures, this given method recognized the attacks from a huge set of the rules that are describing each type of known attack. And major demerit of this approach is the complexity for recognizing the unknown types of attacks. Intrusions within the computing area are the very common unwanted malignant activities which are found while monitoring the resources of computer. Various security approaches have been implemented since the last few decades, but as the Technology is raising so the security threats have also been increasing. Along with the entire world now a days are based on the computers, either directly or may be indirectly, so it is very significant problem to avoid these malignant activities and their threats which may affect the computing architecture [7, 9]. An ID is a software or hardware which can detect any unauthorized use of the system from the internal or external users. Based on the NIST definition, intrusion detection is the process of monitoring the events that took place in the network or system. These logs should be analyzed to detect the intrusions [5]. In this research paper a Modified random sampling method is proposed for NIDS. This paper is organized various topics such as related work in IDS, Problem statement, proposed solution MRFA, implementation, result analysis and in the last conclusions of the entire work with future scope are discussed [22].

2. RELATED WORK IN IDS DETECTION

IDS are a device or software application that monitors network and or system activities for malicious activities or policy violations and produces reports to the management station or administrator [7]. Intrusion detection is defined as the operation of detecting the unwanted traffic over the network or within a device. Figure 2.1 shows IDS system [5].



2.1 EXISTING IDS DETECTION METHOD-

A number of classification techniques are available in weka tool. Following methods are widely used by various researches for IDS detection [7]-

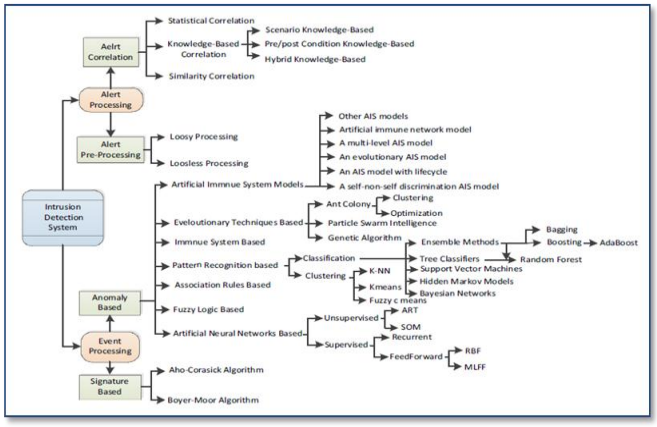


Figure 2.1Exisitng IDS Methods [7]

Naive Bayes classifier-Naive bayes are one of a probabilistic classifiers based on Bayes' theorem with strong i.e. naive assumptions among the features and these assumptions are independent. In 1960 [6], it was described under a name into the text retrieval community [21].

- **Bayes Net classifier**- Bayesian Network [2] is a statistical model that represents a set of r variables which are random and conditional dependencies through a directed graph that is acyclic (DAG).It represents a probabilistic relationship and based on these relationships it finds out the classes of the network traffic coming. Here, a node represents random variables and edges shows conditional dependencies.

- **IBK-** It refers to K-nearest neighbor technique [3]. It is instance based algorithm and when k=1, this means object is simply assigned to single nearest neighbor class.

- **Sequential Minimal Optimization (SMO)-** It is used in SVM (Support vector machine)[4]. It is generally used for solving problems related to quadratic programming. It is implemented by the [2] LibSVM which is a tool used for training of support vector machine. It is an iterative algorithm which picks the multiplier and continues to optimize them until convergence.

- **Random Forest-** It is decision tree based algorithm.[8] It is operated by constructing a multitude of decision trees during training time and output is the class that classify. In terms of intrusion detection, the class is anomaly and normal in which anomaly refers to an attack.

  The benefits of Random Forest are [13]-
  - Random forest can run capably on huge databases.
  - Random forest can handle an N quantity of input data without variable removal.
  - It provides the most essential features in the classification.
  - It can execute well even if the data are omitted.

  The limitations of Random Forest are-
  - The huge number of trees in random forest can be a reason for interruption in processing.
  - It has been noted that random forest is apt only for a few datasets.

- **J48-** It is C4.5 decision tree based algorithm [18]. It is developed as an extension of ID3 algorithm of Ross Quinlan. It is also referred as a statistical classifier and It has ranking #1 in top 10 Data mining algorithms [7]. It is an open source java implementation algorithm available in weka.

- **Random projection-**The random projection [1] technique is used to ease the dimensionality of a group of points. From the name itself, it reveals that it reduces the amount of random variables. Thus the difficulty of organization of large datasets can be reduced. In random projection technique, the original d-dimensional data is projected to k-dimensional (k<<d) subspace all the way through origin. The group of points lies in the Euclidean space. Random projection methods are dominant methods known for their effortlessness and less incorrect results compared with other methods.

- **Support vector machine-**Support vector machine [23] is a technique that has been emerged for the analysis of

data for the classification process. The support vector machine is also known as support vector networks. The SVM uses a set of training data where each one has been labeled into one of two categories. The training data set builds a model and the new unknown data would be categorized into the proper group. There will be a linear separation between the data that has to be classified. With the help of this line the data can be easily separated with more accuracy.

2.2 POPULAR DATASET USE IN NIDS-

Following datasets are widely used in NIDS detection system-

2.2.1 KDD-Cup99- This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment [20].

2.2.2 NSL-KDD- It is a data set suggested to solve some of the inherent problems of the KDD'99 data set which are mentioned in [11]. Although, this new version of the KDD data set still suffers from some of the problems discussed by McHugh and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. Furthermore, the number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.

2.2.3 Kyoto Dataset- The Kyoto University Benchmark dataset [19] consists of 3 years (November 2006 through August 2009) of data captured from honey pots, dark net sensors, a mail server, a web crawler, and Windows XP installation. While very carefully constructed and comprehensive, the dataset does not lend itself to evaluation of new NIAD classifiers for several reasons: First, the Kyoto dataset contains only the values of specified features, and lacks the full raw-data packet captures which would allow for implementation of future advances in feature selection and extraction [15].

Data Set	Number of attributes	Sample Traffic Data	Classes Can be Identified
KDD Cup-99	42	More than 100000	5
NSL-KDD	42	More than 100000	2
Kyoto-2006	24	More than 200000	3

Table 2.2 Comparison of data sets [6]

3. PROBLEM STATEMENT-

The main objective of this research is to develop an efficient intrusion detection system by improvement in existing methods. Random forest is a widely used method in NIDS. Random forest is one of the most successful classifiers and ensemble learning. It is a classification algorithm obtained through extending the decision tree classifiers using ensemble learning techniques. The existing original random forest (RF) algorithm has some challenges in feature selection process, selection of classifiers, selection strategy for random features for various training and also challenges in combination steps. In this research paper we are presenting a network intrusion detection system based on Modified Random forest classifiers. Proposed Modified random forest algorithm (MRFA) is combination of unpruned classifiers and CART (regression tree) with bagging approach. From the selected features, MRFA selects the best features and built the decision tree, a sampling variable and confusion matrix are used for identification of data more accurately and efficiently.

Proposed method will achieved following results over existing methods-

- Better correlations in between datasets and classifiers performance.
- Better results for true positive rate, false positive rate, precision and F-measure.
- Efficient IDS detection.

4. PROPOSED SOLUTION MRFA

In this research paper we are presenting a network intrusion detection system based on Modified Random forest classifiers (MRFA).

Algorithm for MRFA-

Input- Data set D with various entries.

($D_{trainging}$= {D1, D2…….Dn} where $D_{trainging}$= training data set

Output- Resulting data set $D_{resulting}$, with better detection dare and accuracy

Step-1 Create decision tree and learning

1.1 Retrieve data set and upload it on weka

Step-2 Tree bagging by bootstrap method

2.1 Apply bagging or bootstrapping approach on the data set, it divides the data set into different subset of data set with replacement of rows
2.2 Given a training set X = x1, ..., xn with responses Y = y1, ..., yn,
2.3 Bagging repeatedly (B times)
2.4 Selects a random sample with replacement of the training set and fits trees to these samples:
For b = 1, ..., B:Sample, with replacement,
B training examples from X, Y;
2.5 call these Xb, Yb.
2.6 Train a decision or regression tree fb on Xb, Yb.
2.7 After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':-

$$F' = \frac{1}{B}(x+a)^n = \sum_{k=1}^{B}(fb)x'$$

2.8 Select by taking the majority vote in the case of decision trees.

Step-3 from Bagging to Random Forest

3.1 Check all the value in each field, column and for each attribute
3.2 Find the best selection
3.2.1 Applying Splitting_criteria ( )
3.2.2 Generates all the label nodes based on splitting criteria function
3.3 Apply CART ()
3.3.1 Measure accuracy for each each of the Decision Tree

Step-4 Extra Trees

4.1 Check the entire extra tree
4.2 Match with new results

Step-5 Merging

5.1 Attach a leaf labeled with the majority class in D to node N
5.2 Attach the node return by generate decision tree (generate () ;) to node N;
Step-6 Return Tree_N

5. IMPLEMENTATION AND RESULT ANALYSIS-

Proposed MRFA method is implemented by using Java and simulated on weka tool and tested on Kdd-Cup 99 and NSL-Kdd dataset, and compared with Naive b ayes, J-48 and Random forest. Various performance measuring parameters such as true positive rate, false positive rate, precision and F-measure are calculated.

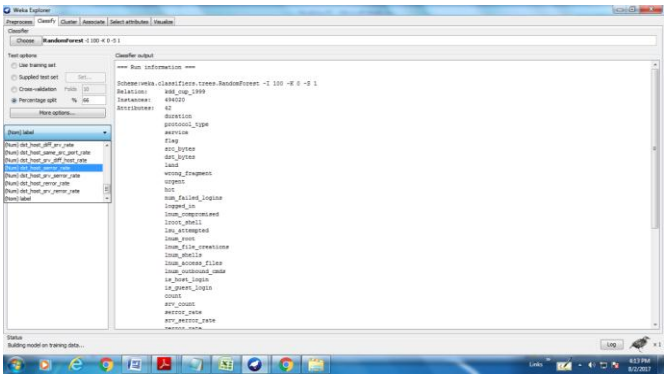


Figure 5.1 Weka (Implementation of proposed method)

5.1 Comparison Parameters-

Following parameters are used to compare existing and proposed method-

5.1.1 Accuracy- This refers to the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

Accuracy= TP + TN / TP+ FP+ TN + FN

Where – True positive (TP) – classifying normal class as normal class, True negative (TN) – classifying anomaly class as anomaly class. , False positive (FP) – classifying normal class as an anomaly class and
False negative (FN) – classifying anomaly class as a normal class

5.1.2 Detection ratio- It is defined as the ratio of detecting attacks to total no of attacks. This is the best parameter to measure the performance of the model. Detection ratio means correctness in a model for detecting intrusion.

Detection ratio = $TP/TP+FN$

5.1.3 True positive ratio- This is one in which correct classification of data has been performed. Means correctness in a system to detect normal or abnormal data. It is defined as-

TPR= TP / FN+ TP

5.1.4 False positive ratio- This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. It is defined as- $FPR = FP/TN+FP$

5.1.5 Precision ratio- It is also known as Positive Predictive Value (PPV). It measures the relevant instance that is retrieved after classification. High precision means that classifiers or algorithm returns more relevant results.

PPV= TP/ TP+FP

5.1.6 Recall- It is also known as sensitivity. This is also used to measure the relevant instance, which is selected. The higher value of recall more the relevant data is selected for classification. It is defined as- **Recall = TP / TP+ FN**

5.1.7 F-Measure- It is basically used to measure the effectiveness of the classifiers. This is harmonic mean of precision and recall. It is also known as traditional F-measure or balanced F-score. It is defined as-

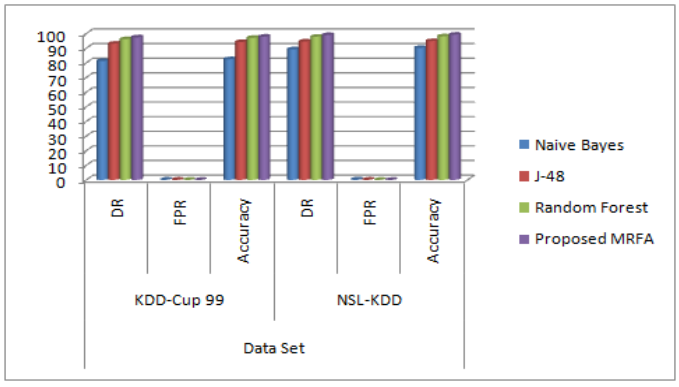
F_measure= 2* {(Precision * Recall) / (Precision + Recall) }

5.2 Result Analysis for KDD cup 99 & NSL-KDD-

For Kdd-cup 99 and NSL-KDD following results were calculated-

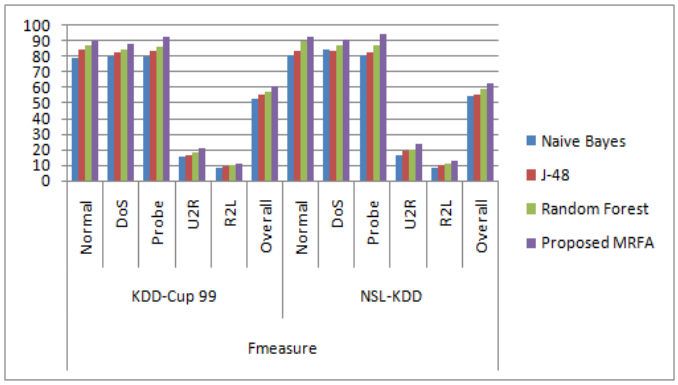
5.2.1Detection Ratio (DR) & false positive Ratio (FPR) & Accuracy (in %)-

Method	Data Set					
	KDD-Cup 99			NSL-KDD		
	DR	FPR	Accuracy	DR	FPR	Accuracy
Naive Bayes	81.25	0.25	82.5	89.26	0.33	90.21
J-48	93.11	0.23	94.25	94.56	0.21	94.85
Random Forest	96.15	0.12	97.1	97.8	0.09	98.1
Proposed MRFA	97.4	0.1	97.9	99.01	0.07	99.25



5.2.2 F-measure %- Following result are calculated for F measure- Below table 5.2.2 show result for F measure %, for proposed method and existing methods.

Method	F-measure											
	KDD-Cup 99						NSL-KDD					
Class Name	Normal	DoS	Probe	U2R	R2L	Overall	Normal	DoS	Probe	U2R	R2L	Overall
Naive Bayes	78.8	80.25	79.98	15.6	8.56	52.64	81.25	84.56	80.52	16.58	8.4	54.26
J-48	84.25	82.9	83.65	16.55	9.2	55.31	83.25	83.68	82.56	19.25	9.9	55.73
Random Forest	87.56	84.56	85.96	18.25	9.8	57.23	89.98	86.85	86.96	20.25	11.3	59.07
Proposed MRFA	90.25	88.56	92.35	21.22	11.25	60.73	92.35	91.25	94.56	23.6	12.5	62.85



Influences- The above results and graph clearly shows that proposed MRFA performs outstanding in terms of accuracy, detection rate and F-measure for both data sets (KDD-Cup99 and NSL-Kdd), over existing methods

6. CONCLUSIONS & FUTURE WORKS

In this research paper we are presented a network intrusion detection system method based on Modified Random forest classifiers. Proposed method has some unique features such as sample variable; cast matrix features which over comes the feature of existing Random forest method. An experimental study clearly shows that proposed method MRFA performs outstanding in terms of accuracy, detection rate and F-measure for both data sets (KDD-Cup99 and NSL-Kdd), over existing methods such as Naive bayes, J-48 and Random forest.

In future work we can modify selection policies and criteria by creating new rules. We can also implement our proposed method with more realistic data sets and can compare with more methods.

REFERENCES-

1. Nabila Farnaaz∗ and M. A. Jabbar,"Random Forest Modeling for Network Intrusion Detection System", Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016). ScienceDirect, Procedia Computer Science 89 (2016)PP 213-217

2. Lata, InduKashyap, '' Study and Analysis of Network based Intrusion Detection System'', International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013

3. N.S.CHANDOLIKAR, V.D.NANDAVADEKAR, ''Comparative Analysis of two Algorithms for intrusion attack classification using KDDCUP Data Set'', International Journal of Computer Science and Engineering (IJCSE) Vol.1, Issue 1 Aug 2012.

4. Devikrishna K, Ramakrishna B, ''An Artificial Neural Network based Intrusion Detection System and

Classification of Attacks '', International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015

5. Vaishali Kosamkar, Sangita S Chaudhari,'' Improved Intrusion Detection System using C4.5 Decision Tree and Support Vector Machine'', International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014.

6. Rajendra V. Boppana, Senior Member, IEEE, and Xu Su, Member, IEEE A Distributed ID for Ad Hoc Networks, 26th International Conference on Advanced Information Networking and Applications 2012.

7. Leila Mechtri, Fatiha Djemili Tolba, Salim Ghanemi, MASID,"Multi agent based intrusion detection in MANET", IEEE 2012.

8. Monita waghengbam and ningrila marchang,"Intrusion detection in MANET using fuzzy logic", IEEE 2012.

9. X. Zhang, L. Jia, H. Shi, Z. Tang and X. Wang, "The Application of Machine Learning Methods to Intrusion Detection", Engineering and Technology (S-CET), 2012 Spring Congress on, (2012), pp. 1-4.

10. Robert Mitchell and Ing-Ray Chen, "Behavior Rule Specification-based Intrusion Detection for Safety Critical Medical Cyber Physical Systems", IEEE Transactions on Dependable and Secure Computing (Volume:12 , Issue: 1), 2014

11. Devikrishna K S and Ramakrishna B B, "An Artificial Neural Network based Intrusion Detection System and Classification of Attacks", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, Issue 4, Jul-Aug 2013, pp. 1959-1964

12. Vaishali Kosamkar, Sangita S Chaudhari, "Improved Intrusion Detection System using C4.5 Decision Tree and Support Vector Machine", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1463-1467

13. Nilofer Shoaib Khan, Prof. Umesh Lilhore, "Review of various intrusion detection methods for training data sets", International Journal of Modern Trends in Engineering and Research (IJMTER)
Volume 03, Issue 12, December – 2016, PP 197-202

14. Nilofer Shoaib Khan , Prof. Umesh Lilhore, "An Efficient NIDS by using Hybrid Classifiers Decision Tree & Decision Rules", International Journal of Science & Engineering Development Research (IJSDR), Volume 2 Issue 1, January-2017, PP 76-79

15. Jamal Hussain, Samuel Lalmuanawma, Feature analysis, evaluation and comparisons of classification algorithms based on noisy intrusion dataset, ScienceDirect,2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016), PP 188 – 198

16. Safaa O. Al-mamory , Firas S. Jassim, On the designing of two grains levels network intrusion detection system, ScienceDirect, Karbala International Journal of Modern Science 1 (2015) 15e25

17. G.V. Nadiammai, M. Hemalatha, Effective approach toward Intrusion Detection System Using data mining techniques, Egyptian Informatics Journal, Elsevier, 2013, PP 37-50.

18. Ghorbani AA, Lu W, Tavallaee M. Network Intrusion Detection and Prevention Concepts and Techniques, Springer New York Dordrecht Heidelberg London; 2010.

19. Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications 2014; 41(4): 1690–1700

20. Shwarz SS, Singer Y, Srebro N. Pegasos: Piramal estimated sub-gradient solver for SVM. In: Proceedings of the 24th International conference on machine learning; 2007. p. 807-814.

21. Freund Y, Schapire RE. Large margin classification using the perceptron algorithm.Machine Learning 1999; 37(3): 277-296.

22. Khardon R, Wachman G. Noise tolerant variants of the perceptron algorithm. The Journal of Machine Learning Research 2007; 8: 227-248.

23. Wettschereck D, Dietterich T. Improving the performance of radial basis function networks by learning center locations. In Neural Information Processing Systems 4. Denver, CO: Morgan Kaufmann NIPS 1992; 4: 1133–1140.