A Comparative Analysis of Various Clustering Techniques for Sales Fraud Detection

B. Kharthik Kumar Reddy[1], N.Ch. Sriman Narayana Iyengar[2]*

Abstract: Nowadays, sales fraud increasingly becomes common in our society. In this regard, sales fraud detection must be required to prevent such schemes in every organization. This paper deals with the implementation of various clustering techniques such as k-means, k-modes, hierarchical clustering, partitioning around medoids (PAM), and also the self-organizing map (SOM) techniques to efficiently analyze and detect the presence of sales fraud. The results of the comparative analysis state that the self-organizing maps can be used efficiently to detect sales fraud as compared with the other considered clustering techniques.

Keywords: k-means, k-modes, hierarchical clustering, medoids, Self-organizing maps, Sales fraud detection

1. Introduction

Sales play a vital role in the growth of any organization. It helps in building trust and long-term adherence between customers and businesses. The organizations' sales department is responsible for building strategic decisions to make their business grow. They are also responsible for the pricing of products based on customer satisfaction to increase sales to a greater extent. They must keep track of the record of the pricing details of all the goods or services, product information, and its distribution. Customer feedback during sales transactions can be essentially valuable for the improvement of products and services.

The occurrences of sales fraud were due to unwanted manipulation of the records related to the sale of goods or services or giving of false sales information. In addition, intentional manipulation of the actual details on business profit or loss and expenditures in a particular period is also considered as signs of fraud. Fraud can occur in the purchasing department, sales department, or account department. All departments in the organization or business must maintain a related ledger to maintain their record which can be used for the analysis of profits or losses.

[1] Information Technology Department, Sreenidhi Institute of Science & Technology, Hyderabad, Telangana, India
Email: bkkreddy@gmail.com
[2]* Information Technology Department, Sreenidhi Institute of Science & Technology, Hyderabad, Telangana, India
Email: srimannarayanach@sreenidhi.edu.in (Corresponding Author)

Received [July 26, 2020]; Revised [October 20, 2020]; Accepted [November 10, 2020]

Published by InnoCon Publishing
ISSN 2704-4440

Sales fraud can be done either through physically stealing goods or products or by giving false information regarding the details of the products or services. The records on both cases were manipulated which greatly affects the organization while taking the next strategic decision for their growth. Sales fraud can be based on the type of goods being stolen which were easily reflected on the records of the organization. Generally, sales fraud can be accomplished by providing false sale information and false purchase information. A sales transaction record includes information on the selling of goods and on the money that comes in the organization or business. False sales information occurs whenever the record only provides the distribution of goods and there was no information regarding the flow of money. False purchase information is the opposite of false sales information wherein the supplier gets paid during sales but product delivery may not be satisfactory. The inventory may have recorded that the payment was already deducted, but there are no product details for its delivery. This false information can be provided by the supplier or employee of the organization or business.

Cash sales can result to void sales since the correct details of sales can be altered or hidden from the records. For example, products sold to customers paid by cash and were not issued with receipts can result in the loss of product details from the company records but actually, it has been delivered to the customer which has not been recorded.

A false return is a fraud related to the refund of money to the customer. There are instances that products can be returned by customers but the refund of payments is not within their policies. There are also instances that a refund is done but with excessive deduction.

Most businesses use barcodes to record their product details. Whenever a product is sold, barcodes are read to show the product details and price. During sales, if a certain product is not initialized with a barcode, then the sales record for such a product does not exist. This kind of fraud directly involves the organization for hiding product sales records so that they can increase their sales without showing it to the government.

Fraud risk analysis refers to the process of determining the probability of the fraud occurrence, the fraud detection techniques, and the necessary steps to prevent it to occur. Cost-effective action plans for fraud risk analysis can be analyzed which can be beneficial to organizations. The analysis of frauds can be implemented through the following actions:

- Threat identification. There are four main possible threats involved in any organization, namely, the financial, informational, operational, and strategic threats. The financial threats are considered to be the main cause of fraud occurrences, money losses, loss of financial information, and expenditures in unnecessary activities.
- Estimation of the risks and losses from threats. The probability that risk can occur and the reoccurrence of the loss of data will be estimated. In addition, the involved losses will be determined.
- Identification of the prevention steps. After the loss identification and risk estimation, proper decisions must be taken to prevent the occurrence of threats and the way to deduct the fraud that can occur must be outlined. The prevention of the occurrences of fraud is better than its deduction.
- Cost Analysis. The cost to control the existence of fraud must be determined. Effective fraud analysis is necessary before it can be controlled and prevent its occurrence. Necessary prevention steps must be taken according to fraud risk levels as there’s no thumb rule for risk analysis in an organization. Cost analysis in fraud prevention and control is essentially beneficial to the organization.

This paper deals with the comparative analysis of the implementation of various clustering techniques for sales fraud detection. The clustering techniques include the k-means clustering, k-modes clustering, hierarchical clustering, partitioning around medoids, and the self-organizing map techniques. It aims to find the best technique to efficiently analyze and detect the presence of a sales fraud.

The rest of this paper is organized as follows: Section 2 discusses the review of related literature; Section 3 presents the model for sales fraud detection; the various clustering techniques for sales fraud detection were outlined in Section 4; and the concluding remarks are presented in Section 5.

2. Review of Related Literature

A fraud detection framework for the use of credit cards was created using pattern recognition and signal processing where the number of fraudulent transactions is lesser than the legitimate transactions [1]. The likelihood score ratios are combined together to provide a better fraud detection solution. Surrogates can be created and pre-processed from real data and inputted to various classifiers for training and testing to produce scores. The scores are then combined to estimate the key performance indicators (KPI) which will be used to get the results. The fraud detection techniques used were Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Non-Gaussian Mixture analysis (NGM), Random Forest analysis (RF), and Support Vector Machine (SVM).

Data has been extracted from transactions by aggregating the transaction to observe and analyze the spending behavior of customers [2]. The von Mises distribution was used to create different sets of features by analyzing periodic customer behaviors. It used various sets of features on the real credit card dataset of a European credit card company in order to analyze their impact. The results of using the proposed periodic features showed an average increase in savings by 13 percent.

A fraud detection system that addresses the limitations of the existing fraud detection systems such as scalability issues, extreme imbalanced class, and time constraints was presented by Mareeswari and Gunasekaran [3]. This can be done through the utilization of a commonly used method for pattern recognition and classification referred to as the hybrid support vector machine (HSVM) along with communal and spike detection. The extracted raw data were converted into the required form and given as input to the HSVM classifier.

The Predictive Analysis Technologies (PAT) was used by Hafiz *et al*. [4], wherein, their capabilities, relevant criteria, and features were evaluated to prepare a scorecard. The various PAT vendors considered were Falcon Fraud Manager, IBM SPSS Manager, SAS Fraud Manager, Cyber Source Decision Manager, and ACI Proactive Risk Manager. They have been evaluated based on the relevant features and criteria to prepare the scorecard.

The Hidden Markov Model (HMM) was used by Bhusari and Patil [5] to detect frauds on credit cards that have obtained high fraud coverage and low false alarm rate. The implementation of HMM has created clusters that depict the cardholder's spending details. The transactions are classified into three clusters based on whether the amount spent is low, medium, or high.

An Artificial Neural Network (ANN) technique was proposed by Srivastava et al. [6] to be used in checking whether the transactions were fraudulent at payment gateways. The data in the pavement gateways were combined with the data provided by the merchant. The extracted data were then normalized to generate rules which are then provided to the classifier. The transactions were divided among four categories such as fraudulent, doubtful, suspicious, and not fraudulent.

A credit card fraud detection that utilizes an Artificial Neural Network (ANN) and the meta-cost method was presented by Ghobadi and Rohani [7]. It is then observed that the rate of fraud detection has been increased and the cost has been decreased with the use of such methods. The process starts with the dataset training using neural networks and each record will be relabeled. Each record was relabeled through passing a number of neural networks and the results will be averaged in order to find the probability of getting a genuine or fraud transaction. The relabeled records were provided as input to the ANN classifier to detect whether a particular transaction is fraudulent or not.

Big data technologies such as Hadoop, HBase, and Spark were used to detect frauds from the considered dataset wherein large amounts of data were processed in detecting frauds in real-time [8]. Various classifiers were combined to improve fraud prediction accuracy. There are four layers involved that include data storage, batch processing, key values sharing, and streamlines detection.

Modi *et al*. [9] states that credit card fraud detection is becoming essentially important as the usage of credit cards rapidly increases. Various algorithms were applied such as clustering algorithm, prediction algorithm, and forecasting algorithms. The technique for fraud analysis presented by the Authors was completely based on the user's card transactions and reducing false-positive transactions.

3. Sales Fraud Detection Model

Sales fraud is one of the major challenges faced by every organization. It is capable of providing millions or even billions of profit losses to various companies. Companies and organizations are challenged in building a cost-effective risk analysis tool to identify and minimize fraud activities in their sales transactions. This section presents a sales fraud detection model that can adapt to the behavioral changes of fraud activities. The model will consider various clustering and machine learning algorithms to effectively determine or identify fraudulent sales transactions from genuine sales transactions.

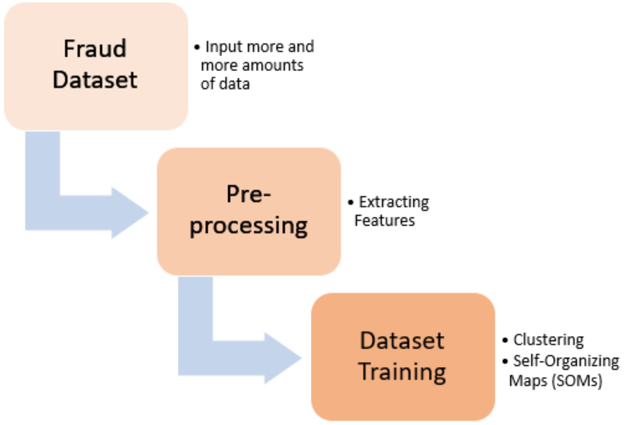

Figure 1. The Sales Fraud Detection Model

Figure 1 outlines the fraud detection model which is comprised of three phases. The first phase is the determination of the fraud dataset and fed into the model. The greater amount of data being trained results in greater performance of the fraud detection model. The second phase involves pre-processing where information of each and every attribute associated with a sales transaction process will be extracted. These can include the identity of the customer, identity of the sales personnel, product information, mode of payments, the network used for the transaction, location where the transaction has been made, and others.

The third phase involves the training of the algorithm. Transactions data will be provided to the fraud detection model in order for the algorithm to determine and identify fraud sales transactions from genuine sales transactions.

4. Comparative Analysis of Various Clustering Techniques for Fraud Detection

This section provides an analysis of the different clustering techniques as depicted in Figure 2 that includes k-means clustering, k-modes clustering, hierarchical clustering, and partitioning around medoids (PAM). In addition, the Self-organizing maps (SOMs) learning method was also introduced for comparison with the identified clustering techniques.

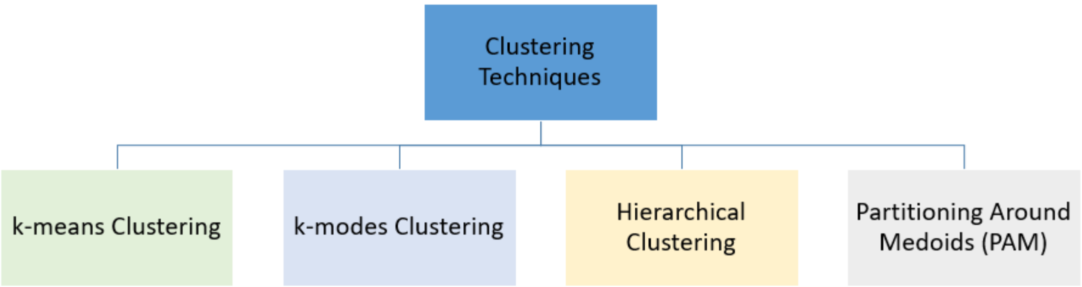

Figure 2. Clustering Techniques

Clustering is a technique that divides data into different groups or clusters. The concept of clustering is that the intra-cluster similarity of data must be high and the inter-cluster similarity must be less (*i.e.*, the data points of different clusters should be dissimilar).

4.1 k-means Clustering

The k-means clustering algorithm [10] is a widely used unsupervised learning algorithm. It is a method of vector quantization that aims to partition *n* observations (data points) into *k* clusters. Each data point belongs to a particular cluster with the nearest mean which serves as the prototype of the cluster.

The distance between the data point and the cluster centroid is calculated in Equation 1.

$$d(cluster\,centroid) = \sqrt{(x-x_c)^2 + (y-y_c)^2} \tag{1}$$

Table 1. The k-means Clustering Algorithm

Phases	Process
Step 1	Initialize the *k* cluster centroids representing the number of clusters in consideration.
Step 2	Calculate the distance of each point from the centroids initialized and place the data point in the cluster such that the distance of the data point from the centroid of the cluster is minimum.
Step 3	Calculate the new centroids of the clusters formed.
Step 4	Repeat steps 2 and 3 until the centroids don’t change, *i.e.*, they converge.

4.2 k-modes Clustering

The k-modes clustering technique [11] has been widely used in situations that involve categorical data. The statistical inference measure such as mean cannot be calculated for categorical data, and k-modes exactly handle this limitation of the k-means algorithm. A dissimilarity measure like hamming distance in Equation 2 can be used for the data involving categorical data.

$$d(x, y) = \sum_{i=1}^{n} (Z_i, Q_i) \tag{2}$$

Table 2. The k-modes Clustering Algorithm

Phases	Process
Step 1	Initialize *k* clusters by choosing the *k* initial modes that may or may not be present in the dataset in consideration.

Step 2	Calculate the similarity of the object with respect to the cluster modes available by the dissimilarity measure as:
	if(object!=cluster_mode) then dissimilarity=1
	if(object==cluster_mode) then dissimilariy=1-nrj/total

Step 3	Place the data point considered into that cluster for which the dissimilarity value is minimum.
Step 4	Recalculate the cluster modes based on the new cluster objects added.
Step 5	Repeat Steps 2 to 4 until the cluster modes remain the same, *i.e.*, they converge.

4.3 Hierarchical Clustering

Hierarchical Clustering [12] is a technique that aims to divide the given data points into a hierarchy of clusters. Hierarchical clustering can either be done through the “Agglomerative method of hierarchical clustering” or the “divisive method of hierarchical clustering”. The Agglomerative method of hierarchical clustering referred to as the bottom-up approach involves placing each data point or the data object into a cluster of its own. Then, based on the similarity of the objects, their respective clusters are merged to form a bigger cluster. This process is repeated until all of the similar objects are placed into a single cluster.

On the other hand, the divisive method of hierarchical clustering referred to as the top-down method involves placing all the data points or the data objects into a single cluster. Then, based on the dissimilarity of the data points, the data objects are placed into different clusters. This process is repeated until all the data points that are dissimilar are placed into different clusters. The hierarchical clustering algorithm is depicted in Table 3.

Table 3. The Hierarchical Clustering Algorithm

Phases	Process
Step 1	Place all the data points into their own respective clusters. The distance between the clusters is equal to the distances between the data points they contain.
Step 2	Find a set of two clusters such that the distance between them is minimum and put them into the same cluster by merging the two respective clusters.
Step 3	Calculate the distance between the new and the old cluster.
Step 4	Repeat Steps 2 and 3 until all the data points are located within a single cluster.

The different hierarchical clustering techniques involve the following:

- *Single linkage*: Represents the shortest distance between a data point in one cluster and the data point from the other cluster.
- *Complete Linkage*: Represents the largest distance between a data point in one cluster and the data point from the other cluster.
- *Average Linkage*: Represents the average distances of a data point in one cluster to all the data points in the other cluster.
- *Centroid*: Represents the distance of the data points that denote the centroids of the clusters taken into consideration.
- *Ward*: Represents the sum of squares of the data points of the two clusters.

4.4 Partitioning Around Medoids (PAM)

Since the k-means clustering algorithm is sensitive to the outliers, a better solution is provided by the partitioning around medoids (PAM) clustering technique [13]. In this algorithm, the clusters are not represented by the centroids, but by the medoids. This algorithm aims to minimize the sum of dissimilarities between the objects in a cluster and the center of the same cluster. The PAM algorithm supports all datatypes.

Table 4. The Partitioning Around Medoids (PAM) Algorithm

Phases	**Process**
Step 1	*k* medoids were randomly chosen to represent *k* clusters.
Step 2	The distance of the data points from the cluster medoids will be calculated.
Step 3	Assign the data point to that cluster that is closest to the medoid of the cluster under consideration.
Step 4	Add the distances of the data points from the medoids to get the total distance.
Step 5	Select a point that is not a medoid and swap it with the current medoid.
Step 6	Reassign every data point to the closest medoid’s cluster.
Step 7	Calculate the total cost.
Step 8	If the calculated total is less than the total distance then keep the new point as the next medoid.
Step 9	Repeat Steps 5 to 8 until all the medoids converge.

4.5 Self-Organizing Map (SOM)

A self-organizing map (SOM) which is also known as a self-organizing feature map (SOFM) is a competitive learning method and considered the most widely used artificial neural network (ANN) technique [14]. SOM is trained using unsupervised learning and does not need human interventions in order to learn and also has no knowledge regarding the input data properties. SOM produces a low-dimensional, discretized representation of the input space of the training samples, called a map. SOM does not need the labels for the input data since it provides a mapping from higher dimensions to the map units that preserve the topology. The points that are nearer to each other are placed as adjacent map units. The nearby map units form a lattice and thus mapping is from a high dimension to a plane. The SOM also preserves the distance between the data points under consideration.

The update formula for a neuron v with weight vector $W_v(s)$ is given by Equation 3 [14],

$$Wv(s+1) = Wv(s) + \theta(u,v,s) * \alpha(s) * (D(t) - Wv(s)) \tag{3}$$

where,

s is the step index,

t an index into the training sample,

u is the index of the best matching unit (BMU) for the input vector $D(t)$,

$\alpha(s)$ is a monotonically decreasing learning coefficient;

$\theta(u, v, s)$ is the neighborhood function which gives the distance between the neuron u and the neuron v in step s.

Table 5. The Self-organizing Map (SOM) Algorithm

Phases	Process
Step 1	Initialize the map by choosing random values for the initial weights.
Step 2	Perform sampling by choosing a set of vectors from the input space.
Step 3	Choose the neuron whose weight is closest to that of the chosen vector.
Step 4	Apply the update formula for a neuron.
Step 5	Repeat steps 2 to 4 until the SOM becomes constant or does not change.

4.6 Comparative Analyses

The dataset used in the analyses of sales fraud detection consists of five attributes as shown in Table 6.

Table 6. Sales Fraud Detection Attributes

Attributes	Description
ID	Represents the identification of a sales personnel.
Prod	Represents the identification of the product.
Val	Represents the reported transaction value.
Quant	Represents the quantity of a particular product.
Insp	Contains a report with three possible values (OK, Fraud, and Unkn).

The dataset presented is in categorical form. The attributes were pre-processed and converted the categories, namely, OK, Fraud and Unkn, so that the data can be effectively applied to various classifiers (*i.e.*, k-means clustering, k-modes clustering, hierarchical clustering, PAM, and SOM) to be used. “OK” is encoded as ‘0’ which indicates that there is no fraud, “Unkn” is encoded as ‘1’ which indicates that the status as to whether the transaction is fraud or not is unknown, and “Fraud” is encoded as ‘2’ which means that fraud is detected in the sales transaction.

This dataset was then applied to the different clustering techniques that were identified and discussed. The observations on the results of the algorithms were shown in Table 7.

Table 7. Comparison of Clustering Techniques

Clustering Technique	Accuracy (%)
k-means Clustering	94.5
k-modes Clustering	94.8
Hierarchical Clustering	98.8
Partitioning around Medoids (PAM)	94.8
Self-organizing Map (SOM)	99.2

It is observed that the self-organizing maps give the highest accuracy of 99.2%. Hierarchical clustering also gives good results of 98.8%. The k-means clustering algorithm obtains an accuracy of 94.5%, and the k-modes clustering algorithm and the PAM algorithm obtained an accuracy of 94.8%

5. Conclusion

This paper deals with the analysis of the different clustering techniques for sales fraud detection such as the k-means clustering, k-modes clustering, hierarchical clustering, partitioning around medoids (PAM), and the self-organizing map (SOM) technique. Based on the results of the analyses, the SOM method outperforms the other clustering techniques by obtaining 99.2% of accuracy, followed by the hierarchical clustering with an accuracy value of 98.8%. Therefore, it can be concluded that the SOM method can be efficiently and effectively used in the detection of sales fraud.

In the future, the convergence of the different clustering techniques will be considered to deliver a more robust algorithm in detecting sales frauds.

References

[1] A. Salazar, G. Safont, A. Rodriguez, L. Vergara, “*Combination of Multiple Detectors for Credit Card Fraud Detection*”, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), December 12-14 2016, Limassol, Cyprus, pp.212-217, doi: 10.1109/ISSPIT.2016.7886023.

[2] A. C. Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, “*Detecting Credit Card Fraud using Periodic Features*”, IEEE 14th International Conference on Machine Learning and Applications (ICMLA), December 9-11,2015, Miami, FL, USA, pp.208-213, IEEE, doi: 10.1109/ICMLA.2015.28.

[3] V. Mareeswari, G. Gunasekaran, “*Prevention of Credit Card Fraud Detection based on HSVM*”, 2016 International Conference on Information Communication and Embedded Systems (ICICES), February 25-26, 2016, Chennai, India, IEEE, pp.1-4, doi: 10.1109/ICICES.2016.7518889.

[4] K. T. Hafiz, S. Aghili, P. Zavarsky, “*The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada*”, 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), June 15-18, 2016, Las Palmas, Spain, IEEE, pp.1-6, doi: 10.1109/CISTI.2016.7521522.

[5] V. Bhusari, S. Patil, “*Study of Hidden Markov Model in credit card fraudulent detection*”, 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), IEEE, February 29 - March 1, 2016, Coimbatore, India, pp.1-4, doi: 10.1109/STARTUP.2016.7583942.

[6] A. Srivastava, M. Yadav, S. Basu, S. Salunkhe, M. Shabad, "*Credit card fraud detection at merchant side using neural networks*", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), March 16-18, 2016, New Delhi, India, IEEE, pp.667-670.

[7] F. Ghobadi, M. Rohani, "*Cost sensitive modeling of credit card fraud using neural network strategy*", 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), December 14-15, 2016, Tehran, Iran, IEEE, pp.1-5, doi: 10.1109/ICSPIS.2016.7869880.

[8] Y. Dai, J. Yan, X. Tang, H. Zhao, M. Guo, "*Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies*", 2016 IEEE Trustcom/BigDataSE/ISPA, August 23-26, 2016, Tianjin, China, pp. 1644-1651, doi: 10.1109/TrustCom.2016.0253.

[9] H. Modi, S. Lakhani, N. Patel, V. Patel, "*Fraud Detection in Credit Card System Using Web Mining*", International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, no. 2, April 2013, pp. 175-179.

[10] I. Dabbura, "*K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*", towardsdatascience.com, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a#:~:text (accessed April 12, 2020).

[11] A. Chaturvedi, P. Green, J. Caroll, (2001). "*K-modes Clustering*", Journal of Classification, vol. 18, no. 1, January 2001, pp.35-55, doi: 10.1007/s00357-001-0004-3.

[12] C. R. Patlolla, "*Understanding the concept of Hierarchical Clustering Technique*", towardsdatascience.com, https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec (accessed April 12, 2020).

[13] XLSTAT, "*Partitioning Around Medoids*", xlstat.com, https://www.xlstat.com/en/solutions/features/partitioning-around-medoids (accessed April 12, 2020).

[14] T. Kohonen, T. Honkela, (2011). "*Kohonen network*", Scholarpedia, vol. 2, no. 1, 2007, pp.1568, doi: 10.4249/scholarpedia.1568.