XIX IMEKO World Congress
Fundamental and Applied Metrology
September 6–11, 2009, Lisbon, Portugal

IMPORTANCE OF SCALING IN UNSUPERVISED DISTANCE-BASED ANOMALY DETECTION

Pekka Kumpulainen [1], *Mikko Kylväjä* [2], *Kimmo Hätönen* [3]

[1] Research scientist, Automation science and engineering, Tampere University of Technology, Tampere, Finland, pekka.kumpulainen@tut.fi
[2] Senior Solution Architect, Aditro, Helsinki, Finland, mikko.kylvaja@aditro.com
[3] Senior specialist: Security research, Nokia Siemens Networks, Research, Technology and Platforms, Espoo, Finland, kimmo.hatonen@nsn.com

Abstract – One of the key applications in mobile network monitoring is to detect anomalous phenomena in the network. Distance-based methods are commonly used in unsupervised anomaly detection. The results are dependent on the distance metrics used and the scaling of the variables. In many cases very simple methods can provide sufficient performance if the variables have been scaled properly. In this paper we discuss the importance of scaling in distance-based methods and the possibility to incorporate a priori knowledge of the relative importance of the variables by scaling. We present an example of a priori scaling on performance data measured from the radio interface in a mobile telecommunication network. The results are compared to those obtained by using traditional normalization.

Keywords: anomaly detection, scaling, a priori information

1. INTRODUCTION

Detection of anomalies or outliers is an important task in many data analysis applications. The amount of data collected from industrial applications is ever-increasing and it is impossible for the process operators to browse all the data manually. Therefore automated methods are required to detect the samples or parts of the data that contain something that might be of further interest.

Anomalies are not necessarily always signs of errors or malfunctions. They can also reveal new valuable information from the system [1]: *An apparently wild (or otherwise anomalous) observation is a signal that says: ''Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study"*.

A general definition for an outlier was given by Hawkins [2]: *An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. This definition is very extensive but it gives no guidelines how to determine whether an observation is an outlier or not.

Various statistical methods have been used in outlier detection [3]. In the usual case in mobile network monitoring, as well as in many other industrial applications, there are no data with predefined labelled anomalies available. Thus supervised methods can not be used to detect anomalies. Therefore unsupervised methods have to be used. The normal or most common behaviour of the system can be modelled from the history data. The states that are rare and deviate from the common behaviour can be detected as anomalies. Detecting anomalies in single variables is simple enough: they are located at the tails of the distribution. While nonlinear transformations do change the differences between samples, their order in the distribution will remain as long as the transformations are monotonic. Anomaly detection in multivariate data is more sensitive to transforms. Even linear scaling of the variables will affect the order in which the anomalies are ranked according to their severity.

Simplicity is usually preferred in industrial applications. Complexity of the methods also increases the complexity of the application they are used in, therefore increasing the requirements for maintenance. In many cases very simple distance-based methods in anomaly detection can provide sufficient performance provided that the variables have been scaled properly.

Unfortunately the scaling, also referred as weighting, of the variables is often neglected and underrated part in research. Normalization to zero mean and unit variance is accepted as a standard procedure without further investigations whether it is the most suitable method for the case or not.

In this paper we discuss the effects of scaling on unsupervised distance-based anomaly detection. We present a scaling method that is solely based on a priori expert knowledge. We demonstrate the method using two radio interface performance measurements from a mobile telecommunication network. The results are visualized and compared to those obtained by normalization which is a very common method for scaling. Finally we present an example using four variables and demonstrate the effect of scaling on the detected anomalies.

ISBN 978-963-88410-0-1

2. UNSUPERVISED ANOMALY DETECTION

Unsupervised anomaly detection is a wide field. The methods vary from principal components [4] to self organizing neural networks [5] and clustering [6, 7]. Local features of the data space can also be emphasized [8, 9]. Knorr et. al. [10] have defined a distance-based outlier. However, also other previously mentioned methods rely on distance metrics. In industrial applications it is important that the anomalies can be ordered according to how critical states the anomalous samples present. That allows the operators to react to the most severe situations first.

2.1. Distance metrics

Proximity or similarity of two points on the data space is determined by the distance between them [11].

Distance between two points a and b is a function $d(a,b)$. Distance metric has to satisfy three conditions [12] specified in Table 1.

Table 1. Properties required from distance metric.

Property	Definition
positivity	$d(a,b) \geq 0$; $d(a,b) = 0$ iff $a = b$
symmetry	$d(a,b) = d(b,a)$
triangle inequality	$d(a,c) + d(c,b) \geq d(a,b)$

The most common metric is Euclidean distance, also known as straight line distance. If the points a and b are represented by vectors $\mathbf{x}_a$ and $\mathbf{x}_b$ then Euclidean distance is

$$d(a,b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^T} \quad (1)$$

A more general metrics that also takes into account the covariances of the variables is the Mahalanobis distance

$$d(a,b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)\mathbf{S}^{-1}(\mathbf{x}_a - \mathbf{x}_b)^T} \quad (2)$$

where $\mathbf{S}$ is the sample covariance matrix. Mahalanobis distance is scale invariant in a sense that any linear transform $\mathbf{x}' = a*\mathbf{x} + b$, where a and b are scalar constants, does not affect the distance metric. However, nonlinear transformations do affect it and the relative importance of the variables is completely determined by the covariance matrix.

3. SCALING

Scaling plays a very important role in all the methods related to proximities or distances between samples. This applies to clustering [13] as well as anomaly detection [9]. Scaling defines what kind of anomalies will be detected [14, 15]. Proper scaling should make the variables equal in their importance within the problem in which they are used. Decisions about the relative importance of the variables always require process knowledge and experience concerning the variables. This makes it more difficult to implement applications and to adapt to new processes. However, according to our experience utilizing expert knowledge will improve the performance of methods in practice.

Scaling by the range (dividing the data with the range it covers and typically shifted to the range from 0 to 1) has been found to give the best results in clustering [16, 17]. That is, however, very sensitive to outliers in the data. In many real life applications, especially in telecommunication networks with non-normal distributions, more robust and possibly nonlinear methods are required, such as logarithm transforms and robust standard deviations [9]. In this paper we concentrate to traditional normalization and a special piece-vice linear scaling that utilizes a priori knowledge of the importance of the variables.

3.1. Normalization

Normalization, also called z-score or autoscaling, is a very common procedure for scaling. Each variable is forced to have zero mean and unit variance by subtracting the sample mean and then dividing by the sample standard deviation (3).

$$z = \frac{x - \bar{x}}{s_x} \quad (3)$$

Normalization is used very often. It is easy to apply and appropriate when the variables have different scales originally.

Figure 1. shows an example of such a case. Scatter plot of heights in meters and weights in kilograms of 11 persons is presented on original scale. Two interpoint distances are also depicted: distance from sample A to B is 2 and the distance between A and C is 0.17. It is clear that this is not reasonable considering the data set.

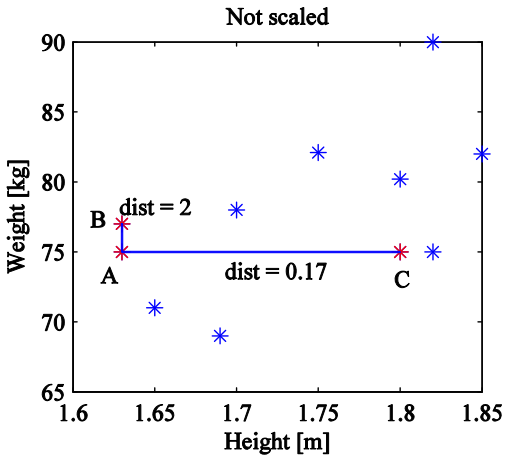

Fig. 1. Scatter plot of heights and weights of 11 persons on original scale.

The same data are presented in Fig 2. Now the variables are normalized. The plot looks the same but now the x and y axes are normalized units. The distance between A and B is now 0.34 and the distance between A and C equals 2.05. This seems much more appropriate interpretation of the distances regarding the data set at hand.

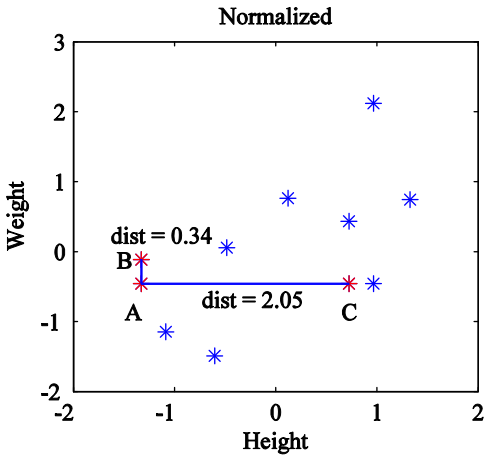

Fig. 2. Scatter plot of heights and weights of 11 persons on normalized scale.

The relative importance of the variables is determined by the variances of the variables. The importance can be adjusted by multiplying each normalized variable by a weight factor corresponding to its relative importance.

In the following section a more general method is presented, where also nonlinear features can be utilized.

3.2. A priori scaling of GSM network performance measurements

The following example shows how a priori information can be integrated in the analysis by a piecewise linear scaling. It allows integration of the end users' a priori knowledge about the behaviour and importance of the variables.

The a priori scaling information is for network performance data of a commercial European GSM operator [18]. All the KPI (Key Performance Indicator) variables selected for analysis have common features: they have a limited range and one end of the range is desired and the other is a major failure. The used variables are network performance indicators that have the range from 0% to 100% and depending if it is a success or failure type of indicator, the desired value is either 100% or 0%.

The network elements' quality indicators are scaled continuously and piecewise linearly to interval [0, 1] extremes corresponding to the worst and the best performance, respectively. The mapping was constructed on the basis of a priori information of network experts. Four values of the performance indicators are defined by experts: *worst possible*, *very poor*, *satisfactory*, and *best possible*. These values are scaled to 0, 0.2, 0.9 and 1 respectively and the scaling function is created with linear interpolation. The scaling function parameters can be adjusted to different performance indicators, different networks and target performance levels. After the scaling all performance indicators are within the same range and the same value refers to the same level of performance in each indicator.

Examples of the piecewise linear scaling of are shown in Fig. 3. The KPIs presented are: *Dropped Call Ratio* (*DCR*) *Radio Down-Link Quality* (*RX_DLQ*) and *Call Setup Success Rate* (*CSSR*). These KPI variables are among the most important metrics in radio network performance analysis.

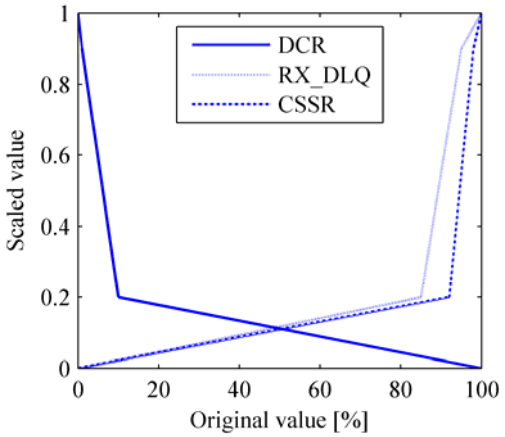

Fig. 3. Example of piecewise linear scaling functions on three variables that all have original scale in percentages.

DCR is an example of a *failure type* variable that has the *best possible* value at 0% (no dropped calls) and the *worst possible* value at 100% (all the calls dropped). *RX_DLQ* and *CSSR* are examples of a *success type* variable. They both have the *worst possible* value at 0% and the *best possible* at 100%. The *very poor* value for *RX_DLQ* is at 85% and for *CSSR* at 92%. The *satisfactory* values are at 95% and 98% respectively.

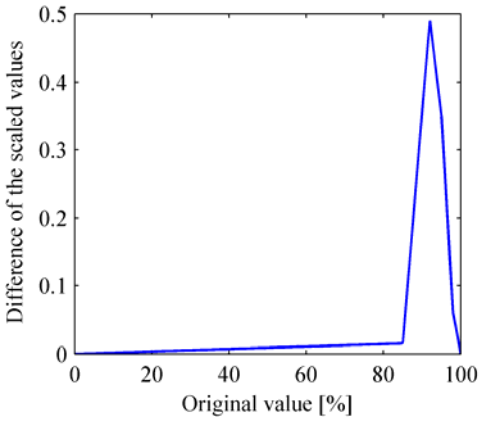

Fig. 4. Difference of the scaling functions of *RX_DLQ* and *CSSR*.

Difference of the scaling functions (Fig. 3) is shown in Fig. 4. They are mostly identical except in the range from 85% to 98%. The maximum difference is at 92% and it is 0.45, meaning 49% of the whole range (0 to 1) in scaled space.

3.3 Examples with two variables

The ideal normal situation in the scaled space will always be ones for every variable. In this application the samples located farthest away from the ideal are considered anomalies. In the actual application there are several

variables but since two dimensions are the most that can be easily visualized, the following examples use two variables: *RX_DLQ* and *CSSR*.

Distance from the ideal normal situation is colour coded in Fig. 5. In the ideal state both variables are at 100% on the upper right corner. White colour equals maximum distance and black equals zero. The samples A and B are located at (90, 95) and (95, 90). The distance from the ideal is the same for both points: 11.18 in the original units.

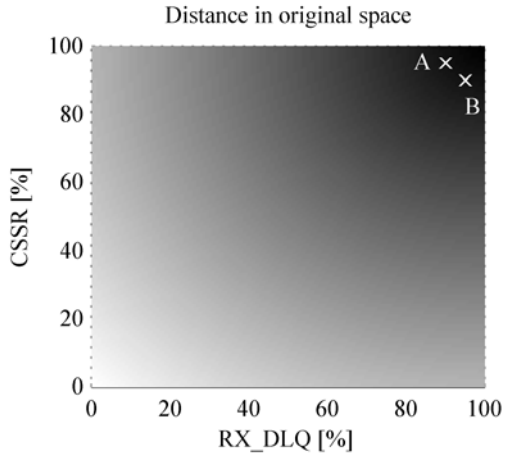

Fig. 5. Distance from the ideal state (both variables 100%). White colour equals maximum distance and black equals zero.

In Fig. 6. the distance is calculated in the scaled space. The colouring presents the distance from the ideal state to that in Fig. 5. The fall of the value of *CSSR* is considered more severe and this has been utilized by the scaling. Therefore the distance of point B from the ideal state is larger than the distance from point A.

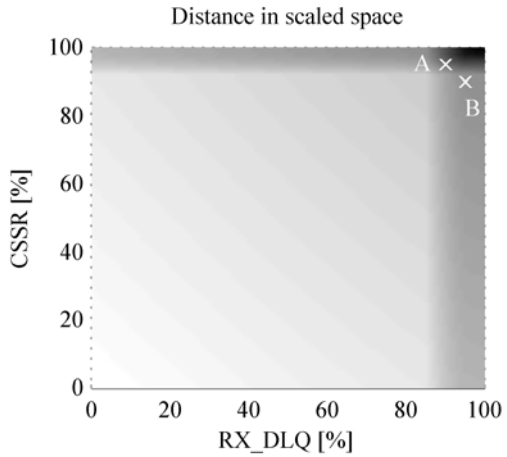

Fig. 6. Distance from the ideal state in scaled space. White colour equals maximum distance and black equals zero.

In the ideal state in scaled space both variables have value 1. The distances from the ideal state are 0.64 for sample A and 0.81 for sample B. The nonlinear piecewise scaling has the effect that sample B is now considered more anomalous than sample A, which is preferred by the end users in this application.

4. EXAMPLES WITH NETWORK DATA

The following example uses the data from micro basestations in a radio network. Basestation is considered micro if it has antenna below rooftop height. There are 191 basestations and the data covers daily performance measurement values for 42 days. With some missing values there are a total of 7752 daily samples. Scatter plot of the data is shown in Fig. 7.

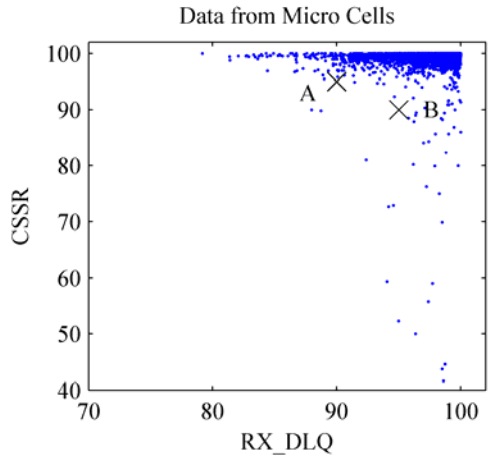

Fig. 7. Scatter plot of the data from Micro cells in original scale. Note the different scales on x and y axis.

Normalization requires a reference data set for calculating the mean and standard deviation. The same example points A and B as previously are also displayed. In the normalized space it is not reasonable to calculate distances from the mean value of the data. Instead, the ideal state is also normalized using the mean and standard deviation of the data set.

The distances of points A and B from the normalized ideal state equal 5.33 and 5.81 respectively. Sample B has greater distance as preferred, but the values are relatively close to each other. The difference of the severity of the samples is not as clear as it is when the piecewise linear a priori scaling was used.

Fig. 8 presents a scatter plot of the a priori scaled data. The ideal state in scaled space is located at [1 1]. The circles present the equal distance contour from the ideal state. The proportions of the data outside the contours are 0.5%, 1% and 2% corresponding to 38, 77 and 155 samples in this data set. Two most severe anomalies (the ones that have the greatest distances from the ideal state) are highlighted with red stars.

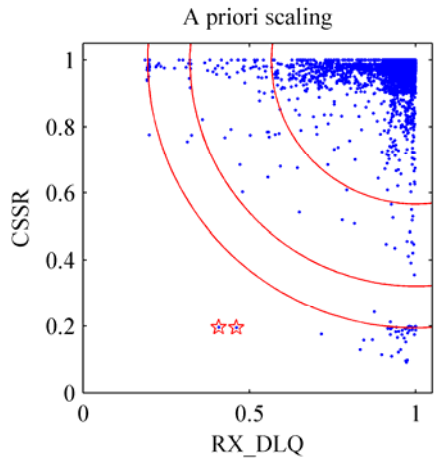

Fig. 8. Scatter plot of the data from Micro cells in a priori scaled space with circles limiting the anomalies.

In addition to plain detection, the anomalies can be clustered to further summarize the information they contain [18]. There is a dense cluster visible in lower right corner in Fig 8 and another one, not so dense in the upper left.

Fig. 9 presents a scatter plot of the normalized data. Using the mean and standard deviation calculated from these data, the ideal state in scaled space is located at [0.95 0.3]. The contours for anomaly thresholds are centred at the ideal state and represent the same percentages as in Fig 8.

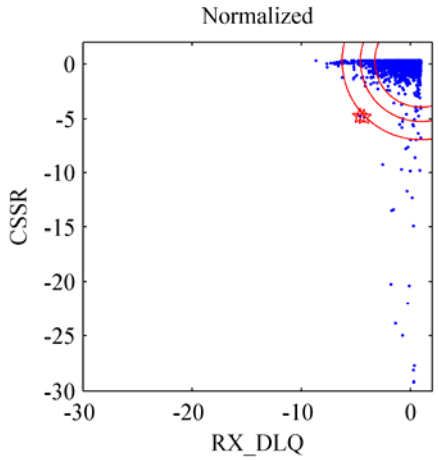

Fig. 9. Scatter plot of the data from Micro cells in normalized scale with circles limiting the anomalies.

The same two samples as above are highlighted with stars. Now in normalized data their positions are 32 and 36 in the order and barely make it to the worst 0.5% samples.

In the normalized space the lower values of *CSSR* are scattered widely and there are no clear clusters visible. This exemplifies well what Gnanadesikan et al. have pointed out [16]: *When done efficiently, weighting and selection can dramatically facilitate cluster recovery. When not, unfortunately, even obvious cluster structure can be easily missed.*

4.1. Anomaly detection using 4 variables

Here we present an example of anomaly detection using four performance variables. In addition to *RX_DLQ* and *CSSR* we have a third success type variable, *Hand-Over Success* (*HO_SUCC*) and one failure type variable *Dropped Call Rate* (*DCR*), which was introduced in Fig. 3. These constitute a typical minimum variable set for network performance monitoring. In four dimensional space single scatter plots can't be used to visualize the whole space. We introduce the effects of scaling using the severity ordering of the anomalies and histograms of the distances from the ideal state.

The additional variables introduced changed the ranking of the anomalies. The two samples highlighted with stars from previous example are now ranked 45 and 47 in priori scaled case, and in normalized case 83 and 88.

The ordering of the top ten anomalies detected using the a priori scaling is compared to the results from the normalized data. Four of the samples are contained in the top ten in both cases. The eighth and tenth samples in a priori scaling are ranked to 53 and 36 in normalized case.

Table 2. Comparison of the order of the most severe anomalies detected using a priori scaling and normalization.

A priori	1	2	3	4	5	6	7	8	9	10
Normalized	5	2	15	4	14	17	1	53	19	36

Fig. 10 shows the histogram of the distances of the samples from the ideal state. All the four performance variables were used and scaled by the piece-vice a priori method. There is a peak in the histogram where the distance is about 0.8. This suggests that there is a concentrated cluster present.

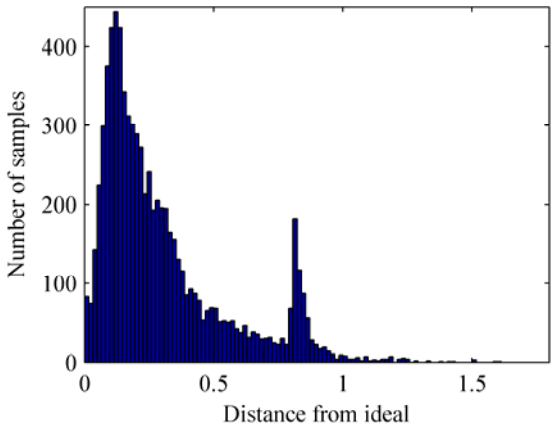

Fig. 10. Histogram of the distances from the ideal state in a priori scaled space.

Similar histogram of the distances from the ideal state with the normalized data is depicted in Fig. 11.

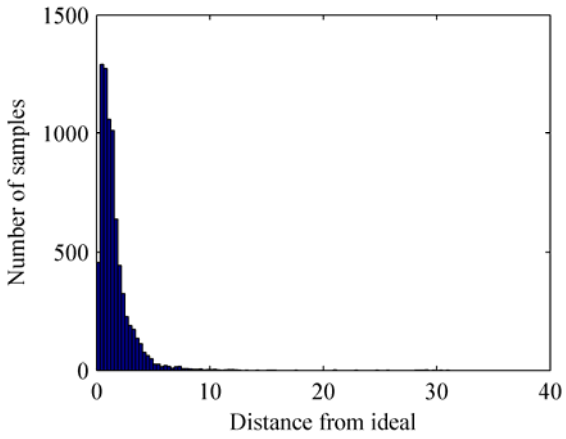

Fig. 11. Histogram of the distances from the ideal state in normalized space.

The distribution of the distances is concentrated close to zero. There are no additional peaks visible and thus the possible cluster structure is hidden in this case. Note that the distances in figures 10 and 11 are not directly comparable, but only the shapes of the distributions are of importance.

5. CONCLUSIONS

Unsupervised anomaly detection is a very wide area of research and application. The range of methods varies from very simple to extremely complicated ones. One key factor is the scaling of the variables. In multivariate data the results of variance and distance-based methods are greatly affected by the scaling. Normalization is widely used but in many cases the results could be enhanced by incorporating a priori knowledge of the process in the methods.

In this paper we have given some simple examples of the importance of scaling. We introduced examples of a piecewise linear scaling method which allows easy integration of end users' knowledge. Such scaling can enhance the performance of distance-based anomaly detection methods.

In order to use the end users' expert knowledge, it has to be acquired from the users. This does not require a reference data set for the scaling and thus it is robust to the imperfection of the data available. It reduces the bias introduced by samples included in the reference data. However, this information is case specific and it has to be tuned for each environment.

In real life applications, like the one presented here, it is unfortunately very difficult to compare the results of various methods as well as the effect of the scaling methods on the results. There are no right answers; the superiority of the methods is solely based on the subjective assessment by the end user.

REFERENCES

[1] W.H. Kruskal, "Some remarks on wild observations", *Technometrics* 2 (1), pp. 1–3, 1960.

[2] D. Hawkins, *Identification of Outliers*, Chapman & Hall, London, 1980.

[3] V. Barnett, *Outliers in Statistical Data*, Wiley, Chichester, 1987.[4] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, L.-W. Chang, A novel anomaly detection scheme based on principal component classifier, in: *Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop*, 2003.

[5] A.J. Höglund, K. Hätönen, A.S. Sorvari, A computer host-based user anomaly detection system using the self-organizing map, *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, vol. 5, IEEE, 2000, pp. 411–416.

[6] K. Leung, C. Leckie, Unsupervised anomaly detection in network intrusion detection using clusters, in: *Proceedings of the 28th Australasian Computer Science Conference*, Newcastle, NSW, Australia, 2005.

[7] P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005.

[8] S. Breunig, H.-P. Kriegel, R. Ng and J. Sander, "LOF: identifying density-based local outliers", in: W. Chen, J.F. Naughton, P.A. Bernstein (Eds.), *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, Dallas, Texas, USA, May 16–18, 2000.

[9] P. Kumpulainen and K. Hätönen, "Local anomaly detection for mobile network monitoring", *Information Sciences*, Vol. 178, Issue 20, pp 3840-3859, 2008.

[10] E. M. Knorr, R. T. Ng and V. Tucakov, "Distance-Based Outliers: Algorithms and Applications", *The VLDB Journal The International Journal on Very Large Data Bases*, 8(3-4), Springer Berlin / Heidelberg. pp. 237-253, 2000.

[11] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining*. Cambridge, MA., USA: The MIT Press 2001.

[12] W. R. Dillon and M. Goldstein, *Multivariate Analysis, Methods and Applications*, John Wiley & Sons Inc., 1984.

[13] B. Everitt, S. Landau, M. Leese, *Cluster analysis*. Edition: 4, Arnold, 2001.

[14] K. Hätönen, P. Kumpulainen and P. Vehviläinen, "Pre and post-processing for mobile network performance data", in: *Proceedings of seminar days of Finnish Society of Automation*, Helsinki, Finland, September 2003.

[15] K. Hätönen, S. Laine and T. Similä, "Using the LogSig-function to integrate expert knowledge to Self-Organising Map (SOM) based analysis", *IEEE International Workshop on Soft Computing in Industrial Applications*, Birmingham University, New York, June 23–25, 2003.

[16] R. Gnanadesikan, J. Kettenring, S. Tsao, "Weighting and selection of variables for cluster analysis", *Journal of Classification*. Vol. 12, 1, 113-136, 1995.

[17] G. Milligan, M. Cooper, "A Study of Standardization of Variables in Cluster Analysis", *Journal of Classification*, 5, 181-204, 1988.

[18] M. Kylväjä, P. Kumpulainen and K. Hätönen, "Information summarization for network performance management", in: M. Laszlo, J.V. Zsolt (Eds.), *Proceedings of the 10th IMEKO TC10 International Conference on Technical Diagnostics*, pp. 167–172, Budapest, Hungary, 2005.