Risk Assessment Using Local Outlier Factor Algorithm

Božidara Cvetković[1,2], **Mitja Luštrek**[1,2]

[1] Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia

[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

boza.cvetkovic@ijs.si

Abstract. In this paper we introduce the unsupervised machine-learning algorithm named Local Outlier Factor (LOF), for health risk assessment. In general the LOF algorithm is used with numerical attributes and the outcome of the algorithm is parting the patterns into normal and abnormal events. In this paper we introduce the extended LOF algorithm with three experimental contributions: (i) utilization of complex nominal attributes, (ii) the developed methodology for detecting the level of event anomaly (low risk, medium risk and high risk) and (iii) providing the information about the risk status for each analysed parameter.

Keywords: anomaly detection, unsupervised machine-learning, LOF, nominal attributes, health risk assessment

1 Introduction

The purpose of the medical expert systems is to disburden the workload of physicians and ease the detection of abnormal events. Research in this field is quite mature. However, modules that assemble the expert system are based on predefined rules created by an expert or models trained on the labelled data. For example, when a patient's health is normal, the parameters characterizing it usually follow some recurrent patterns. When the patient's health is not normal certain parameters move from the normal state and influence others. The rules and models created to detect the risk are highly correlated with the disease they were created for. This means that in case we would like to analyse a different disease, new domain rules have to be created and models re-trained. For that we would need a relatively large amount of labelled data.

There are four problems we are focused on in this research (i) can we use unlabelled data, (ii) is it possible to consider the individuality of the patient regarding the pattern of vital signs and their influence to each other, (iii) can we detect the level of the abnormality and (iv) is it possible to detect how much do the analysed parameters contribute to the risk?

In this research we have adopted the Local Outlier Factor (LOF) algorithm, since it seems the most appropriate method to detect the abnormal events using unlabelled data and by that keep the individuality of the person. The algorithm was extended with the procedure for abnormality level detection per monitored parameter. The developed algorithm enables the doctor to see which of the monitored parameters contribute to the overall risk at most.

2 The Anomaly Detection for Risk Assessment

When a patient's health is normal, the parameters characterizing it usually follow some recurrent patterns. Such patterns can be learned and when a new pattern – an anomaly – is detected, the doctor is notified. If the doctor judges the new pattern to be normal, he can indicate this to the anomaly detection sub-component, and the sub-component will not consider such a pattern anomalous in the future.

2.1 Local Outlier Factor Algorithm

We use the Local Outlier Factor (LOF) algorithm [1] to detect anomalies. The algorithm compares the density of data instances around a given instance A with the density around A's neighbors. If the former is low compared to the latter, it means that A is relatively isolated – that it is an outlier. Such outliers are considered anomalous. The LOF algorithm computes the so-called LOF value for each instance, which is a measure of how anomalous the instance is.

To use the LOF algorithm for risk assessment, it must be trained on a number of instances consisting of the parameters of a patient when his/her risk is normal. For the purpose of the anomaly detection sub-component, such risk is considered low, even though it may be high in absolute terms. After the training data is processed, the parameters of the algorithm must be set: (1) the number of neighbors to consider, (2) the low threshold, which separates the LOF values corresponding to low risk (green) from those corresponding to medium risk (yellow), and (3) the high threshold, which separates the LOF values corresponding to medium risk from those corresponding to high risk (red). Finally, the algorithm can compute the LOF values of new instances and assess the risk.

2.2 The number of neighbours and thresholds

To evaluate the performance of the LOF algorithm, both normal (low risk) and anomalous (elevated risk) instances are needed. We use the concept of the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate (TPR or sensitivity) vs. the false positive rate (FPR or 1 – specificity) at all possible thresholds. The TPR is the fraction of instances correctly classified as normal among all the truly normal ones. The FPR is the fraction of instances incorrectly classified as normal among all the truly anomalous ones. An example of the ROC curve can be seen in Fig. 1. Curves above the diagonal indicate a beneficial classifier, and curves below the diagonal a misleading one. The area under the ROC curve (AUC) is a threshold-independent measure of the performance of a classifier.

The selection of thresholds is also experimental. We want the low threshold to be such that few anomalous instances are below it. This means that the FPR must be below a maximum value. We want the high threshold to be such that few normal instances are above. This means that 1 – TPR must be below a maximum value. Finally, the instances between the thresholds (yellow) may be normal or abnormal.

2.3 Individual parameters

The LOF algorithm merely computes how anomalous an instance is, while we are also interested in the contribution of the individual parameters to its anomalousness. Therefore we compute per-parameter LOF values, which are done the same way as for the regular LOF values, except that the distances (d and k-distance) are computed only with respect to the parameter of interest.

3 Experiment and Results

The experiment was done on preliminary data. The data consists of the activity and energy expenditure computed by the CHIRON activity monitoring methods, heart rate, and body temperature of five persons during the following scenarios: lying still, sitting still and standing still, sitting doing light activities, walking and standing doing light chores, scrubbing the floor, sweeping, sit-ups and jumping jacks, walking normally, walking quickly, running slowly, running normally, stationary cycling normally, stationary cycling vigorously.

All the recorded data were considered normal. We split each scenario in four parts, using the first and third part for training, and the second and fourth for testing. We also needed anomalous test data, which we generated by replacing the values of a parameter at one time (for example the heart rate during lying) with the values at another time (the heart rate during walking briskly).

We had to devise a distance measure for the activity parameter, since it is nominal and has no "natural" distance. We represented each activity by the vector of attributes used for the activity recognition, averaged over all the instances of the activity in the training data. We then computed the Euclidean distances between each pair of activity vectors, yielding the activity-distance matrix.

Fig. 2 shows the prototype of the risk assessment for patients with cardiac disease. The first panel shows the overall deviation with the risk detected. The second panel represents the values of the instance. Other panels are per-parameter risks. We can observe that the parameter for energy expenditure is in the medium risk level, shown on the last panel. This indicates that the energy expenditure level is too low for the measured heart beat and the activity.

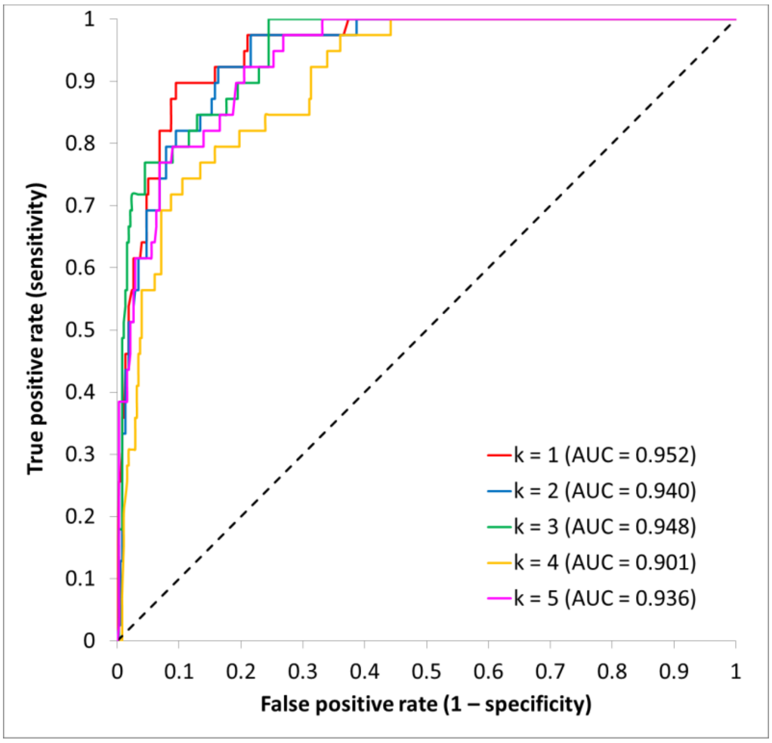
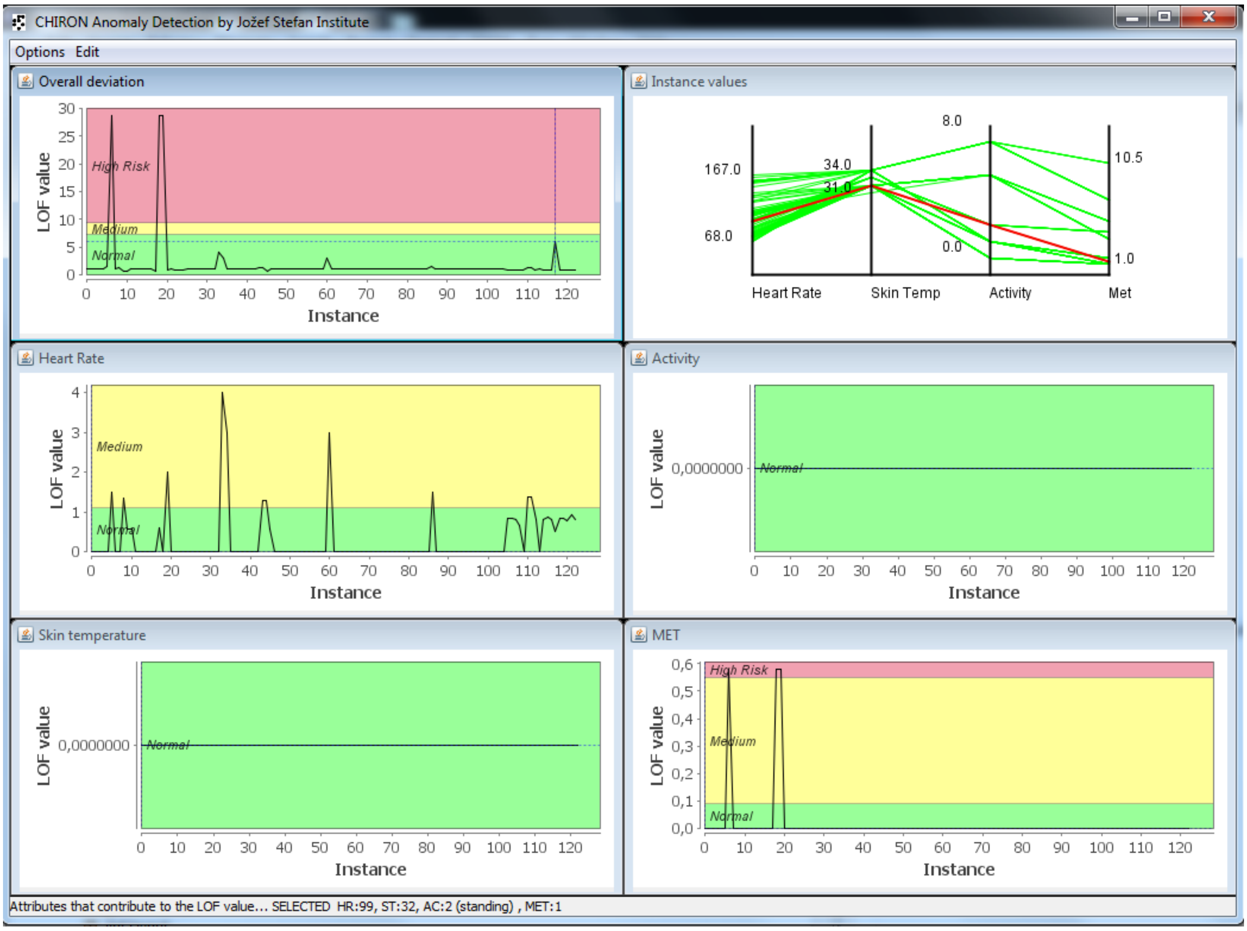


Figure 1: ROC curves for different number of neighbours k = 1, 2, 3, 4, 5.

4 Conclusion

In this paper we have shown that LOF can be used for health risk assessment. We have extended the general LOF to use nominal values in our case activity and to show the level of abnormality.

The disadvantage of LOF as a risk assessment method is that a new pattern is not necessarily a sign of increased risk. However, the advantage is that it can detect any kind of anomaly – there is no need for an expert to describe the possible anomalies and no need for examples of the anomalies (labelled data).



Attributes that contribute to the LOF value... SELECTED HR:99, ST:32, AC:2 (standing) , MET:1

Figure 2: The prototype showing the anomaly detection due to the MET value.

References:

[1] M. M. Breunig,, H.-P. Kriegel, R. T. Ng & J. Sander. LOF: Identifying density-based local outliers. In *Proc. of ACM SIGMOD International Conference on Management of Data,* 2000.

The purpose of the medical expert systems is to disburden the workload of physicians and ease the detection of abnormal events. Research in this field is quite mature. However, modules that assemble the expert system are based on predefined rules created by the expert or models trained on the labelled data. For example, when a patient's health is normal, the parameters characterizing it usually follow some recurrent patterns. When the patient's health is not normal certain parameters move from the normal state and influence others. The rules and models created to detect the risk are highly correlated with the disease they were created for. This means that in case we would like to analyse a different disease, domain new rules have to be created and models trained. For that we would need relatively large amount of relevant labelled data.

There are four problems we are focused on in this research (i) can we use unlabelled data, (ii) is it possible to consider individuality of the patient regarding the pattern of vital signs and their influence to each other, (iii) can we detect the level of the abnormality and (iv) is it possible to detect how much do the analysed parameters contribute to the risk?

In this research we have adopted the Local Outlier Factor (LOF) algorithm, since it seems the most appropriate method to detect the abnormal events using unlabelled data and by that keep the individuality of the person. The algorithm was extended with the procedure for abnormality level detection per monitored parameter. The developed algorithm enables the doctor to see which of the monitored parameters contribute to the overall risk at most.

The disadvantage of LOF as a risk assessment method is that a new pattern is not necessarily a sign of increased risk. However, the advantage is that it can detect any kind of anomaly – there is no need for an expert to describe the possible anomalies and no need for examples of the anomalies (labelled data).