

AN ALTERNATIVE ARTIFICIAL INTELLIGENCE TECHNIQUE FOR DETECTING OUTLIERS

Abstract

Data are rarely perfect. Whether the problem is data entry errors or rare events. Outliers have two opposing properties. They can be noises that disturb regression and classification task. On the other hand, they can provide valuable information about rare phenomena, which can lead to knowledge discovery. This paper proposes a hybrid algorithm including K Nearest Neighbor and Support Vector Machine (KSVM) that detects outliers by taking the advantages of the two intelligent techniques, Support Vector Machine (SVM) and K Nearest Neighbour (KNN). Also a global efficiency measure introduced to compare different methods. Finally, a comparison between KNN, SVM, and KSVM is conducted using detection rate, accuracy rate, false alarm rate, true negative rate and the proposed global efficiency measure based on benchmark data called Milk data.

Key words: Outliers, outlier detection, KNN, SVM, and global efficiency.

1.Introduction

The problem of the existence of outliers in data is an important problem that has been investigated within diverse knowledge disciplines and application domains. The importance of outlier detection is due to the fact that outliers in data are translated to significant (and often critical) information in a wide variety of application domains. Therefore, outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. Many terms have been used including novelty detection, anomaly detection, noise detection, deviation detection, or exception mining, but all involve a similar process or goal. In addition to many different names, there are many more algorithms for solving them. *According to Barnett and Lewis (1994).*

An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the data set.

• Source of outliers

1. Error: An error refers to a noise-related measurement coming from a faulty sensor. Outliers caused by errors may occur frequently, while outliers caused by events tend to have extremely smaller probability of occurrence.

2. Event: An event is defined as a particular phenomenon that changes the real-world state, e.g., forest fire, air pollution, etc. This kind of outliers usually lasts for a relatively long period of time and changes historical pattern of sensor data, (Fawzy et al., 2013).

There are many methods for detecting outliers. The commonly used methods of outlier detection are:

I. Classical techniques include:

- a. Statistical approach
- b. Clustering approach
- c. Density approach
- d. Depth approach

II. Intelligent techniques include:

- a. Classification approach
- b. Distance approach
- c. Information theory approach
- d. Spectral decomposition approach
- e. Visualization approach
- f. Wavelet approach

Most of the classical techniques deal with univariate variable, but most of the problems statistician faced, the data are multivariate or multi-dimension.

The rest of the article is organized as follows. Section 2 presents the evaluation methods. Section 3 involves the new technique (KSVM), its algorithm; in addition, the results of applying the mentioned methods on benchmark data called Milk data.

2. Evaluation methods

Outlier detection algorithms are typically evaluated using the detection rate, the false alarm rate, and the accuracy rate. In order to define these metrics, let us look at a confusion matrix, shown in the following table. In the outlier detection problem, assuming class "C" as the outlier or the rare class of the interest, and "NC" as a normal (majority) class, there are four possible outcomes when detecting outliers (class "C") - namely true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). Global Efficiency is a general measure

combining three efficiency measures of a given method (accuracy, detection rate, negative true rate), into only one value representing the total efficiency of that method. Global efficiency facilitates the comparison between different methods. . It takes values between 0 and 1. More closer to 1, it indicates more efficiency.

Table (1): Confusion matrix, (Yang et al., 2009)

	Predicted outliers class C	Predicted normal class NC
Actual outliers class C	True Positive (TP)	False Negative (FN)
Actual normal class NC	False Positive (FP)	True Negative (TN)

From the previous table, detection rate, false alarm rate, and accuracy may be defined as follows, (Yang et al., 2009):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Detection Rate = \frac{TP}{TP + FN}$$

$$False Alarm Rate = \frac{FP}{FP + TN}$$

$$True Negative Rate (Specificity) = 1 - False Alarm rate = \frac{TN}{FP + TN}$$

$$Global efficiency = \sqrt[3]{Detection Rate * Accuracy * True Negative Rate}$$

Detection rate gives information about the relative number of correctly identified outliers, while the false alarm rate reports the number of outliers misclassified as normal data records (class NC).

We consider data from Daudin et al. (1988) which consists of 8 measurements on the composition of 86 milk container samples. This data set has been studied in many papers in the outlier detection context (e.g. Atkinson, 1994), starting from Caussinus and Ruiz-Gazen (1993).

Daudin et al. (1988) used the data to illustrate a bootstrap study of the stability of principal components analysis; they found only two large components. But, outliers will be in the original eight-dimensional space, rather than in this reduced space. Caussinus and Ruiz-Gazen (1993) extended this analysis using a generalized principal components analysis in which the covariance matrix is estimated robustly; they identified numerous outliers.

These data contains outliers. The 10 most extreme outliers-in order-observations 2, 41, 1, 44, 12, 14, 74, 13, 15 and 3 were all identified outlying by all 100 forward searches, (Atkinson, 1994).

A classical multivariate analysis using the sample mean and covariance estimate would not detect anything. So, KSVM is used to detect outliers. Especially, it is an intelligent technique.

Let us consider a data set D with p features and n instances. In a supervised classification context, we must know also the classes where each of the instances belongs. It is very common include the classes as the last column of the data matrix. The objective is to detect all the instances that seem to be unusual, these will be the multivariate outliers. By using SPSS software program, we test the data (Milk data) by boxplot test for univariate data. One might think that multivariate outliers can be detected based on the univariate outliers on each feature but as it is shown in the figure (1) this is not true. The instance appearing in variables 1, 2, 7 and 8 (which is sample number 12) is a multivariate outlier but it is not an outlier in each feature. On the other hand, an instance can have values that are outliers in several features but the whole instance might be not a multivariate outlier. There are several methods to detect multivariate outlier.

For example, we want to find out the multivariate outliers in each of the variables of the dataset milk by building box plots. People in the data mining community prefer to rank the instances using multi-criteria measures rather to classify the instances in outliers and non-outliers. Rocke and Woodruff (1996) stated the Mahalanobis distance works well identifying scattered outliers, Mahalanobis (1936). However, in data with clustered outliers the Mahalanobis distance measures does not perform well detecting outliers. Data sets with multiple outliers or clusters of outliers are subject to the *masking* and *swamping* effects.

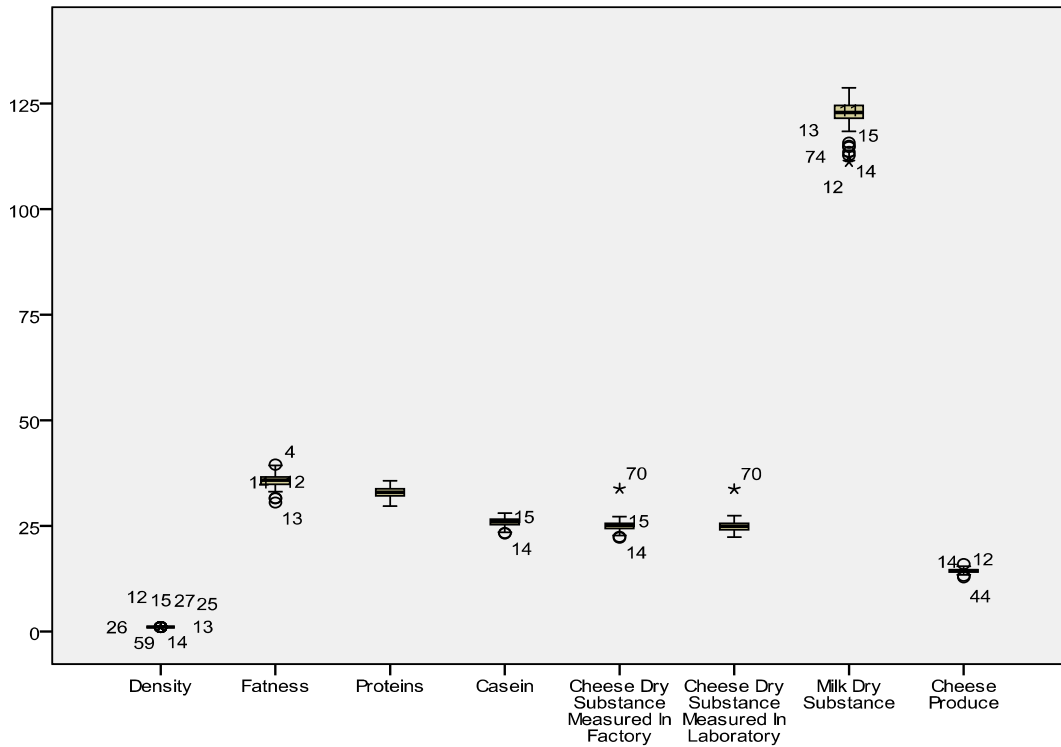


Figure (1): Detecting univariate outliers by box plots in the Milk dataset

- **Masking effect.** It is said that an outlier masks a second one close by if the latter can be considered outlier by himself, but no anymore if it is considered along with the first one. Equivalently, after the deletion of one outlier, other instance may emerge as an outlier. Masking occurs when a group of outlying points skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.
- **Swamping effect.** It is said that an outlier swamps other instance if the latter can be considered outlier only under the presence of the first one. In other words after the deletion of one outlier other outlier may become a “good” instance. Swamping occurs when a group of outlying instances skews the mean and covariance estimates toward it and away from other “good” instances, and the resulting distance from these “good” points to the mean is large making them look like outliers.

3. KSVM method

The KNN approach is outstanding with its simple training and classifying algorithms, due to the fact that KNN implements instant-based learning algorithm. Hence, it does not need construction of a feature space until it receives a command

to classify an input sample. Besides this, it can manage multiclass data set sufficiently by requiring a small amount of training samples. Despite that, most distance metrics, such as that of nearest neighbours, are expected to become less sensitive as the dimensionality of the noisy data increases, thus limiting the performance of NN when applied to data classification. In other words, the performance of KNN classifier is severely degraded by noises. In many cases, KNN is often used as the preliminary classifier to prune the training samples and eliminate those samples effectively to reduce the sample size for better classification, Knorr (1998).

Support Vector Machines (SVMs) were originated from statistical learning theory. The concept of SVMs was introduced by V. N. Vapnik in the late 1970's. SVMs are now regarded as an important example of "kernel methods", arguably the hottest area in machine learning. Machine learning algorithm is the ability of a computer algorithm to recognize patterns that have occurred repeatedly and to improve its performance based on past experiences. SVMs separate different classes by finding optimal classification hyper planes from training, leading to better generalization capability compared with other methods. In SVMs learning, kernel methods are often used to map the data vectors in the input space into a higher dimension feature space thereby converting linearly non-separable case into a separable case. The construction of a linear classification hyper plane in this high dimension feature space is equivalent to a nonlinear decision hyperplane in the input lower dimension space, Vapnik (1995).

Support Vector Machine is a learning algorithm based on the principle of margin maximization. Comparative studies have found SVMs to perform as well as or better than most prevalent learning methods.

Rong et al. were the first who proposed KSVM classifier as an improvement to the accuracy of SVM classifier, (Rong et al., 2002).

Then, (Shen and Lin, 2004) presented a new binary classifier that combines SVM with KNN algorithm for gene expression data classification in order to improve disease diagnosis. Conventional SVM classification produces lower classification accuracy if the sample data of the two classes in a binary classification are all close to the separating hyper plane. When conventional KNN classification has been applied to perform gene expression data classification, KNN distance functions are anticipated to have poor classification performance as the dimensionality of the noisy data increases. The KNN and SVM hybrid classification system proposed in this work, coined as KSVM, is developed due to

the fact that conventional SVM is not able to handle the situation when the training samples are close to the optimal separating hyper plane, (Shen and Lin, 2004).

Zhang et al. proposed an integration of nearest neighbor classifier and support vector machine to classify visual objects, coined as KSVM. The proposed KSVM hybrid classification model can be used for large and multiclass datasets with reasonably lower computational complexity in both training and classifying stage. This property of lower computational complexity is inherited from KNN classification approach, which does not need construction of a feature space, and it is able to handle and manage multiclass datasets with high dimensionality. Due to this fact, the proposed KSVM model is able to provide a more flexible framework from a computer vision perspective as the emphasis is on similarity of objects, (Zhang et al., 2006).

Mei et al. proposed a hybridized algorithm that uses both (KNN) and (SVM) classifiers, coined as KSVM, to analyze gene expression for cancer classification, in both binary and multi-class categorization. They mentioned that precise diagnosis and classification is the ultimate goal for successful treatment of illness, (Mei et al., 2009) .

KNN algorithm has been used by KSVM hybrid classification approach as the initial process of its classification task. In the situation that there is no conflict in the classification of a given input data, where all k neighbours are from a same class, a straightforward classification task is performed by the simple NN algorithm without needing the involvement of SVM approach. On the other hand, if the k neighbours are not with the same class, the hybrid classification approach is used to perform the classification task.

In the proposed KSVM hybrid classification approach, Nearest Neighbor (NN) algorithm is utilized to prune the original training set, so that SVM which acts as a classification engine of this hybrid model does not need to be train with the original training set, but with a pruned training set which insufficient training samples have been eliminated.

This is to ensure a faster training process for SVM with a smaller training set, without sacrificing the sufficiency of classification performance. For the pruned training set, pair-wise distances between the training samples are computed and the distance matrix will be converted to kernel matrix as the training data for SVM. The general idea of this technique is to train the SVM until it preserves the distance function on the set of neighbors and to find close neighbors to an input sample.

SVM is sensitive to these samples intermixed in another class or these boundary samples. The existence of samples in the overlapped region may be harmful to the performance of SVM. The hybridized classifier can improve the precise of classification, so hybridized of KNN classifiers and SVM is proposed to improve the performance of classification.

The first KNN is used to prune training samples and the second KNN is combined with SVM to classify the Milk samples (cases). We first yield the distances matrix which is a symmetrical matrix containing the Euclidean distance between each pair of samples. Then the k nearest neighbors for each sample is sought. In the first KNN, if the class label of training sample is same as the label of the majority of its k nearest neighbors, the training sample is reserved, whereas others are pruned. For the pruned samples set, the second KNN and SVM are applied to classify. If k nearest neighbors has all the same labels, the sample is labeled. Otherwise, SVM will be applied to classify the rest sample. The SVM-KNN steps are described as follows.

Step1: Compute distances of the test to all training examples and pick the nearest K neighbors;

Step2: If the K neighbors have all the same labels, the test is labeled and exit; else, compute the pairwise distances between the K neighbors;

Step3: Convert the distance matrix to a kernel matrix and apply SVM;

Step4: Use the resulting classifier to label the test.

To implement multidimensional SVM in step 3, three variants from the statistics and learning literature have been tried on small number of samples from our data sets. They produce roughly the same quality of classifiers and the SVM is chosen for its better speed.

The new algorithm of KSVM is slow mainly because it has to compute the distances of the test to all training examples. In our setting, this translates into the practice of computing a "crude" distance to prune the list of neighbors before the more costly "accurate" distance computation. The reason is simply that if the crude distance is big enough then it is almost certain that the accurate distance will not be small. This idea works well in the sense that the performance of the classifier is often unaffected whereas the computation is orders-of-magnitude faster.

An additional trick to speed up the algorithm is to cache the pairwise distance matrix in step 2. This follows from the observation that those training examples who participate in the SVM classification lie closely to the decision boundary and are likely to be invoked repeatedly during test time.

So far, there are two perspectives to look at KSVM: it can be viewed as an improvement over NN classifier, or it can be viewed as a model of the discriminative process. From a machine learning perspective, it can also be viewed as a continuum between KNN and SVM: when K is small, the algorithm behaves like a straightforward KNN classifiers. To the other extreme, when $K = n$ our method reduces to an overall SVM.

5.2 Application and results

Applying mentioned illustrated methods (KNN, SVM, and KSVM) on Milk data using Matlab software, the results were shown in table (2). As mentioned before, the data have 8 dimensions and 86 samples, 10 of them are outlier samples. In the training phase 10 normal samples (sample number 4,5,6,7,8,9,10,11,16 and 17) and 5 outlier samples (sample number 1,2,3,12 and 13) are considered. In testing phase, 66 normal samples (sample number from 18 to 40, 42, 43, from 45 to 73 and from 75 to 86) and 5 outlier samples (sample number 14,15,41,44 and 74) are tested by using the three methods mentioned before. The main results are as follows:

- 1- The highest global efficiency score, (0.9148), for detecting outliers is satisfied by the two methods SVM(RBF) and KSVM.
- 2- Although the proposed method KSVM has the same efficiency as SVM(RBF) for the given data. It may be more suitable for other data because KSVM has the advantages of the two methods SVM(RBF) and KNN.

Finally the authors are considering other hybrid between different methods to find the most suitable algorithm having the highest global efficiency.

Table (2): Table of results

	SVM linear	SVM- polynomial	SVM- RBF ($\gamma = 3.5$)	KNN- k=3	SVM-KNN (RBF, $\gamma = 3.5$ & k=3)
Features count	8	8	8	8	8

No. clean training samples	10	10	10	10	10
No. outliers training samples	5	5	5	5	5
No. clean test samples	66	66	66	66	66
No. outliers test samples	5	5	5	5	5
No. true positives (TP)	2	4	4	4	4
No. true negatives (TN)	61	61	65	63	65
No. false positives (FP)	5	5	1	3	1
No. false negatives (FN)	3	1	1	1	1
Accuracy	0.8873	0.9155	0.9718	0.9437	0.9718
Detection rate	0.4	0.8	0.8	0.8	0.8
False Alarm rate	0.0758	0.0758	0.0152	0.0455	0.0152
True Negative rate	0.9242	0.9242	0.9848	0.9545	0.9848
Global Efficiency	0.6897	0.8780	0.9148	0.8965	0.9148

270

271 5.3 Discussion and conclusion:

272 This paper focuses on the outlier detection problem for multivariate data and
273 proposes one measure (global efficiency) to classify different methods according
274 their performance. Conventional outlier detection techniques are not suitable for
275 multivariate data due to the special characteristics and resource limitations of the
276 multivariate data. On the contrary the proposed algorithm is capable to detect

outlier in multivariate data; therefore it offers a relatively more efficient way to detect multivariate data. The proposed approach takes into consideration various characteristics and features of the both methods (KNN and SVM) and evaluated with real life dataset (Milk data). The suggested global efficiency measure shows that our approach can achieve high accuracy rate for identifying outliers, and demonstrate the effectiveness of the proposed approach. In this paper, KSVM technique results are similar to SVM (RBF) results. But, according to the experimental results obtained by Shen and Lin (2004), KSVM classifier has produced a higher accuracy as compared to conventional SVM and conventional KNN. Based on their findings, the better classification performance of KSVM classifier than conventional SVM classifier is due to the fact that more information is carried after the training process is performed, as compared to SVM. Besides, the number of attributes used in the training process will not have high effect to KSVM classifier than the two latter classifiers. In Mei et al.,(2009), KSVM can perform better than SVM only and KNN only, by having the prune function which can eliminate the mislabelled training samples effectively.

Therefore, we can conclude that KSVM technique is a good outlier detector and performs at least as good as other efficient methods. According to the discussion (given before) which shows that KSVM is the best method for classification, consequently it can be also considered the best (Till now) in outlier detection which is a kind of classification.

References:

1. **Atkinson, A.C. (1994)**, "Fast Very Robust Methods for the Detection of Multiple Outliers," Journal of the American Statistical Association, 89, 1329-1339.
2. **Barnett, V. and Lewis, T. (1994)**, "Outliers in statistical data", John Wiley and sons.

- 310 3. **Caussinus, H., and Ruiz-Gazen, A. (1993)**, "Metrics for Finding Typical
311 Structures by Means of Principal Components Analysis," in Data Science
312 and its Applications, eds. J. C. Escoufier and I. Hayashi, Japan: Academy
313 Press.
- 314 4. **Daudin, J.J. Duby, C. and Trecourt, P. (1988)**, "Stability of Principal
315 Component Analysis Studied by the Bootstrap Method", *Statistics* 19, 241–
316 258.
- 317 5. **Fawzy, A. Mokhtar, H.M.O. and Hegazy, O.(2013)**, "Outliers detection
318 and classification in wireless sensor networks", *Egyptian Informatics*
319 *Journal*, pp. 157–164.
- 320
- 321 6. **Knorr, E. M. and Ng, R. T. (1998)**, "Algorithms for mining distance-based
322 outliers in large datasets", In Proceedings of the 24rd International
323 Conference on Very Large Data Bases, Morgan Kaufmann Publishers
324 Inc., 392-403.
- 325 7. **Mahalanobis, P.C., (1936)**, "On the generalised distance in statistics".
326 Proceedings of the National Institute of Sciences of India 2 (1): 49–55.
- 327 8. **Mei, Z., Shen, Q. and Ye, B. (2009)**, "Hybridized KNN and SVM for gene
328 expression data classification" *Life Sci. J.*, 6: 61-66.
- 329 9. **Rocke, D.M. and Woodruff, D.L., (1996)**, "Identification of outlier in
330 multivariate data", *Journal of the American Statistical Association*,
331 91(435):1047–1061.
- 332 10. **Rong, L., Shiwei, Y. and Zhongzhi, S. (2002)**, "SVM-KNN Classifier -
333 A New Method of Improving the Accuracy of SVM Classifier", *Acta*
334 *Electronicasina*, Vo1.30, No.5.
- 335 11. **Shen, X. and Y. Lin, (2004)**, "Gene expression data classification using
336 SVM-KNN classifier" Proceedings of the International Symposium on
337 Intelligent Multimedia, Video and Speech Processing, Oct. 20- 22, Hong
338 Kong, China, pp: 149-152.
- 339 12. **Vapnik, V. N. (1995)**, "The nature of statistical learning theory", Springer-
340 Verlag New York, Inc., New York, NY, USA.
- 341 13. **Yang X, Latecki LJ, Pokrajac D (2009)**, "Outlier detection with globally
342 optimal exemplar-based GMM", In: Proceedings of the 2009 SIAM
343 international conference on data mining (SDM), 145C154.
- 344 **Zhang, H., A.C. Berg, M. Maire and J. Malik, (2006)**, "SVM-KNN:
345 Discriminative nearest neighbor classification for visual category recognition",
346 Proceedings of the IEEE Computer Society Conference on Computer Vision

347 and Pattern Recognition, Volume 2, June 17-22, Washington, DC, USA., pp: 2126-
348 2136.