INT J COMPUT COMMUN, ISSN 1841-9836
8(1):30-36, February, 2013.

Outlier Detection with Nonlinear Projection Pursuit

M. Breaban, H. Luchian

Mihaela Breaban, Henri Luchian
"Alexandru Ioan Cuza" University of Iasi, Romania
E-mail: {pmihaela, hluchian}@infoiasi.ro

Abstract:
The current work proposes and investigates a new method to identify outliers in multivariate numerical data, driving its roots in projection pursuit. Projection pursuit is basically a method to deliver meaningful linear combinations of attributes. The novelty of our approach resides in introducing nonlinear combinations, able to model more complex interactions among attributes. The exponential increase of the search space with the increase of the polynomial degree is tackled with a genetic algorithm that performs monomial selection. Synthetic test cases highlight the benefits of the new approach over classical linear projection pursuit.
Keywords: outlier detection, nonlinear projections, genetic algorithms

1 Introduction

Mining for outliers is of great importance in many domains: fraud detection, disease identification,intrusion detection, fault diagnosis and so on.

Outliers or anomalies represent rare observations/events that deviate from the majority of data either in magnitude or with respect to an overall pattern. Whatever the data, a statistical model can be attached to it and consequently the data can be considered to be generated by a statistical process. In this view, outliers correspond to very low probabilities under the underlying distribution.

Outliers may sometimes simply occur due to erroneous recording of data and not due to meaningful anomalous data which would correspond in the theoretical model to changes in the generative process. Identifying them must be one of the first steps in data analysis as their presence misleads many algorithms from the machine learning area solving tasks like clustering, classification or regression.

This paper proposes and investigates a new framework for outlier detection derived from the classical projection pursuit methodology, aiming at alleviating some of the drawbacks of the standard approach. An analysis of the popular approaches for outlier detection highlight the benefits of the new approach.

The material is structured as follows. Section 2 surveys existing computational methods for outlier detection. Section 3 succinctly describes the framework of projection pursuit. The method we propose is presented in section 4 and is empirically investigated in section 5, while section 6 draws the conclusions.

2 Computational methods for outlier detection

We consider the unsupervised framework of outlier detection: there is no pre-specified generative model for the data and no labels are available to provide examples from which an algorithm could learn.

Several surveys exist in literature that provide a state-of-the-art for outlier detection in this framework [1–4]. Generally, the literature distinguishes among several classes of methods.

Most *statistical methods* for outlier detection are parametric methods: given a certain kind of statistical distribution outliers are detected as those points with low probability of being generated. In the univariate case usually the Normal distribution is used and outliers are considered those observations that lie at a distance larger than k standard deviations from the mean. For the multi-variate case the Mahalanobis distance is considered to compute the distance of the observations from the mean. The main drawback of this approach is that the parameters of the distribution (mean and standard deviations) are computed based on all observations, including the possible outliers, and therefore they may be highly biased. Non-parametric approaches based on standard deviation identify the subset of observations that, after exclusion, determines the highest decrease in variance; using any form of the Minkowski metric the method can be generalized for the multi-variate case. The drawback of this method is that the size of the search space is exponential w.r.t. the number of observations. The first and the third quartiles are also used to identify outliers in the univariate case.

Distance-based approaches [10–13] for outlier detection base their decisions on computing the distances between each data point and its neighbors. These are multi-variate approaches. The main advantage of these methods over statistical approaches is that no hypothesis on the type of distribution is made, hence such methods are applicable without distribution-dependent restrictions. These methods are computationally expensive due to the calculations of distances between data points; therefore, various ways of scaling them up for large databases have been proposed. Still, one important drawback is present: even if they do not make any assumptions on the type of distribution they are not parametric-free methods. The result is very sensitive to some user-tuned parameters like the radius of the neighborhood, the number of neighbors to be used or the threshold indicating the average distance to the neighbors above which a data is considered as outlier.

Clustering-based approaches employ an unsupervised clustering algorithm to identify groups in data. In this context outliers are identified as unusually small groups in data. Popular algorithms used in this context are hierarchical clustering methods (mostly the single-link variant) and density based clustering methods like DBSCAN. The methods have the advantage of being generally applicable to any distributions in data. Their drawbacks reside in high computational costs requiring distance computations between data items and sensitivity to parameters - in case of density-based methods.

All the above-mentioned multi-variate techniques take into account the entire attribute space. Based on distance computations, a drawback is inherent: in high-dimensional spaces the ratio of the distances of the nearest and farthest neighbors to a given target is almost one making outlier detection an impossible task.

Projection pursuit is hardly mentioned in existing surveys on outlier detection. When it is, attention is given usually to a particular exponent of this class of methods - Principal Component Analysis (PCA)- which is in fact a dimensionality reduction method that aims at preserving as much as possible the variance in data and not a method dedicated to outlier detection. Projection pursuit can be used to identify subspaces of the original attribute space where outliers are present, alleviating the mentioned drawback resulting from high dimensionality. This paper highlights the important role projection pursuit can play for outlier identification and enhances the classical methodology by introducing nonlinear projections.

3 Projection pursuit

A k dimensional projection of a data set $X \in R^{n \times d}$ consisting of n items described by d numerical attributes is a linear transformation involving k orthogonal vectors in a d-dimensional space. These vectors form an orthogonal basis $A \in R^{k \times d}$. The projection of X into A is

the product $Z = X \cdot A^T$ resulting in a new representation for each of the n data items in a k-dimensional space.

Projection Pursuit (PP) [5] is a technique aiming at identifying low-dimensional projections of data that reveal interesting structures. The framework of PP is formulated as an optimization problem with the goal of finding projection axes that minimize/maximize a measure of interest called projection index. The projection index can be formulated to identify subspaces where clusters are visible, linear combinations that discriminate between given classes or low-dimensional views of data that reveal the presence of outliers. Depending on the formulation of the index under maximization/minimization analytical methods exist (the case of PCA), gradient-based methods may be used (if the index is continuously differentiable) or probabilistic heuristics like Hill Climbing or Simulated Annealing are employed.

The current work is conducted towards identifying single-dimensional views (one-dimensional projections) of data that present outliers. These can be manually inspected, or simple statistical rules (univariate analysis) can be applied to identify and exclude the outliers.

A popular index used to derive projections with high chances of containing outliers is kurtosis defined as the fourth moment around the mean divided by the square of the variance:

$$kurt = \sum_{l=1}^{n} \frac{(y^{(l)} - \mu)^4}{(n-1)\sigma^4} \tag{1}$$

where μ is the mean and σ is the standard deviation of the single-dimensional projection. A value close to 3 indicates a normal distribution. Higher values indicate the presence of extreme deviations while lower values indicate bimodal distributions. In consequence, this index should be maximized to derive projections containing outliers.

Other indices also exist but are more expensive computationally [6, 7].

One drawback of the classical projection pursuit approach is that linear projections corresponding to linear combinations over the original attributes, are not able to model complex generative models and consequently are not able to detect the outliers in all cases. Such a case is illustrated in Figure 1. To alleviate this drawback we extend the original framework to allow the derivation of nonlinear projections. To this aim, we extend the data set by introducing new features built as products of the original ones. Projection pursuit can be further conducted with classical methods proposed in literature. Because the extension we propose increases considerably the search space, we also design and investigate an optimization framework based on multi-modal genetic algorithms, which allows searching simultaneously for the relevant attributes combinations and the coordinates of the projection axis.

4 An algorithm for detecting outliers with nonlinear projections

Nonlinear projections can be performed after introducing in the analysis new features generated as products of original features. This is a standard approach in other data mining tasks (i.e. regression and classification) to extend standard methods that derive linear models but hardly used in the field of projection pursuit in general and outlier detection in particular. One drawback is inherent to this approach: the exponential increase of the number of new attributes introduced into the analysis with the increase of the degree of the polynomial model.

If the number of original attributes in a data set is m, the number of monomials of degree 2 that can be introduced in the analysis is $\frac{m(m+1)}{2}$ while the number of monomials of degree 3 is $\frac{m(m+1)(m+2)}{6}$. In general, the number of monomials of degree i that can be formed on m variables in $\binom{m+i-1}{i}$ and the number of attributes that can be introduced to derive polynomials up to a certain degree k is $N_k = \sum_{i=1}^{k} \binom{m+i-1}{i} = \frac{(m+k)!}{m!k!} - 1$.

To deal with the large number of features and speed up the projection pursuit algorithm we incorporate a feature/monomial selection mechanism. To this aim a genetic algorithm is designed that simultaneously selects good monomials and searches for good projection axes within the selected monomial subspace.A multi-modal genetic algorithm is used in order to allow for simultaneous exploitation of several good monomial subspaces. A candidate solution corresponding to a chromosome in population consists of two parts:

- a boolean string of length equal to the number of monomials to a given degree k: value 1 corresponds to monomials to be included in search of projections;
- a vector in the Euclidean space playing the role of the projection axis, corresponding in fact to numerical weights in the resulted polynomial transformation

Such mixed representations are common in feature selection tasks solved with genetic algorithms in the context of clustering [15, 16] and classification [14].

The projection $y^{(l)} \in R$ of an item $x^{(l)} \in R^m$ in the subspace encoded by a chromosome is computed as follows: $y^{(l)} = \sum_{i=1}^{N_k} b_i \cdot w_i * m_i^{(l)}$ where $m_i^{(l)}$ is the ith monomial computed as a product over elements from $x^{(l)}$. Using only a subset of the monomials to conduct further the search of axes is equivalent to assigning weight 0 for the rest.

For multi-modal search we use the Multi Niche Crowding GA [8], an algorithm able to maintain stable subpopulations within different niches, to maintain diversity throughout the search and to converge to multiple local optima. MNC is a steady state algorithm that implements replacement based on pairwise comparisons.

Both the selection and replacement operators implement a crowding mechanism. Mating and replacement within members of the same niche are encouraged while allowing at the same time some competition for the population slots among the niches.

Selection for recombination takes place in two steps: one individual is selected randomly from the population; its mate is the most similar individual from a group of size s which consists of randomly chosen individuals from the population. The two chosen individuals are subject to recombination operators and one offspring is created.

The individual to be replaced by the offspring is chosen according to a replacement policy called *worst among most similar*: f groups are created by randomly picking g (crowding group size) individuals per group from the population and one individual from each group that is most similar to the offspring is identified; then, the one with the lowest fitness value among these is replaced. In the original MNC algorithm the replacement is always performed, even if the fitness of the offspring is lower than the fitness of the individual chosen to be replaced. In our implementation we adopt a Simulated Annealing strategy: lower fitness survival is accepted with a probability that decreases during the run of the algorithm.

The similarity between two individuals is computed based on the boolean part of the chromosome that encodes the monomial subspace; the Hamming distance is used.

Recombination between two chosen individuals consists of crossover that generates one offspring which is subsequently mutated.

Dedicated crossover and mutation operators are designed and applied to each of the two segments of a chromosome.

The crossover operator applied to two chromosomes consists in fact of two operations performed independently on the two parts of the chromosomes. Uniform crossover is used on the binary segment encoding the monomial subspace. On the numerical segment, crossover generates each gene of the offspring as a convex combination between the corresponding genes of the parents.

Mutation is also applied in two distinct phases. Each gene in the binary segment is flipped with a given probability which we call binary mutation rate. To each weight in the numerical

segment corresponding to a selected monomial a random value in the interval (-0.25, 0.25) is added with a probability called weights' mutation rate. The binary mutation rate is lower than the weights' mutation rate in order to encourage better exploitation of a given subspace for optimal projection axes.

Before evaluation, the weight vector in the selected subspace of the offspring is normalized to unit length. The evaluation consists in computing the projection of all data items on the axis given by the weight segment of the chromosome in the encoded monomial subspace, followed by the computation of a projection pursuit index dedicated to detecting clusters in data. The index we use is the kurtosis. We choose to maximize this index mainly because its reduced computational complexity compared to other proposed indices for cluster detection. When the projection is normalized to mean 0 and variance 1 the index consists in summing up the values raised to the fourth power, divided by the total number of items. Usually, projection pursuit is preceded by a linear transformation on the data called sphering that guarantees that every linear projection is distributed with mean 0 and standard deviation 1, eliminating the need for further normalization. A sphering procedure is also applicable in our case after all monomials to degree k are added to the original data. However, in our experiments we normalize each projection prior to computing the kurtosis.

Without a multi-modal search scheme, outliers should be identified and eliminated incrementally, based on linear combinations at first, then monomials of degree 2, 3 and of higher order can be introduced iteratively. The multi-modal algorithm allows for several single-dimensional projections maximizing the index to be returned in one run. This brings some advantages: in a standard heuristic only one solution is returned; identifying all outliers requires iteratively the exclusion of the identified outlier and a new execution of the algorithm. The benefits of multi-modal search were recently highlighted in the context of linear PP [9].

5 Experiments

Figure 1 represents a data set containing 100 observations in a two-dimensional space. It is illustrative for the drawback of classical PP: linear projections of data on one axis are not able to identify the interior outliers.

The parameters of the new method are set as follows: s = 0.15 ˇ pop size, g = 0.10 ˇ pop size, f = 0.15 ˇ pop size. The mutation operator is applied at different rates on the two segments of a chromosome: approximately one mutation during 10 iterations is applied on the binary segment while 1 mutation per iteration is applied on the numerical segment; using a steady-state scheme, only one offspring is generated and evaluated at each iteration. The population consists of 50 individuals, randomly initialized: on the binary segment approximately 4 monomials are selected (set to 1) while on the numerical segment the values are generated in the interval [-1,1].

The algorithm was executed at first with monomials of degree one, to simulate one run of a classical PP algorithm: the search is conducted in the original feature space. The black line in Figure 1 a) represents the projection axis generating a linear combination of attributes of maximum kurtosis, returned by our method in this step. Figure 1 b) represents the histogram of the data under this linear combination of maximum kurtosis: the exterior outlier appears at the left while the interior outliers get mixed with the rest of the data.

Without excluding the identified outlier the algorithm was executed again including in the search space all monomials of degree 2. Figure 1 c) represents the distribution of the nonlinear combination (of maximum kurtosis) of the same data derived with our method: the three interior outliers can be identified at the right, outlier "1" being at the extreme, followed at the left by outliers "2" and "3". As a second test case we are interested in the ability of our method to detect the outliers when noise attributes are introduced. To this aim, 3 uniformly-distributed attributes

are added to the data set in Figure 1, the result consisting in 100 observation in a 5-dimensional space. Figure 2 presents the results of two runs of our algorithm: the linear projection (b) is similarly oriented in the original space, with the outlier at the left and the nonlinear projection (b) identified the three interior outliers at the right.

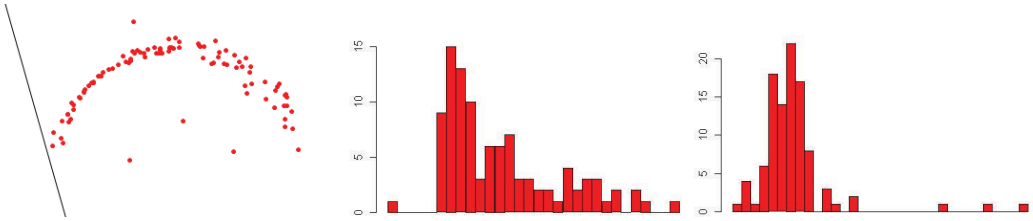

Figure 1: A synthetic data set containing 4 outliers a) The projection axis detected with classical PP is drawn in black; b) The histogram of the linear projection detected with classical PP: only the exterior outlier can be identified; c) The histogram of the nonlinear projection returned with new method: the three interior outliers appear at the right

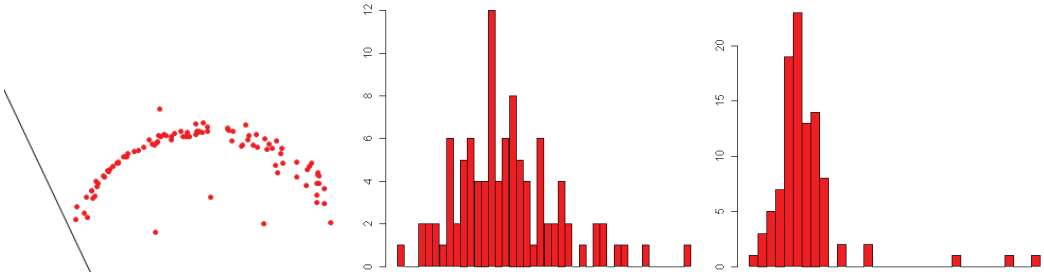

Figure 2: A synthetic data set containing 4 outliers in a 2-dimensional space and 3 more uniformly-distributed attributes a) The projection axis detected with classical PP is drawn in black; b) The histogram of the linear projection detected with classical PP c) The histogram of the nonlinear projection returned by the new method

6 Conclusions

The paper proposes an extension of the classical linear projection pursuit framework in the context of outlier detection. By creating nonlinear projections of data the new method is capable of modeling diverse generative processes and identify outliers which linear projection pursuit does not detect. The proposed method compares positively to distance-based and density-based approaches: the new method provides results which are not altered by the presence of many uniform/gaussian attributes, as it happens with the other approaches, where distance computations over the entire space of attributes are performed. This is because PP intrinsically performs subspaces/attributes selection and moreover, our method deals with monomial selection explicitly. At the same time, the (non)linear combinations of attributes identified to contain outliers can provide useful explanatory information to the user on the nature/source of outliers. Excepting the univariate analysis performed in the last step for outlier exclusion, our method is parameter-free. Scaling up for very large databases is favored by the fact that the proposed algorithm can be easily parallelized: as the most demanding step is projecting the entire data set on a given axis, this operation can be executed in parallel for distinct observations.

Acknowledgement

This work was supported by POSDRU/89/1.5/S/63663 CommScie grant.

Bibliography

[1] H.-P. Kriegel, P. Kröger, A. Zimek, Outlier Detection Techniques, Tutorial at 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, 2010.

[2] V. Hodge, J. Austin, A Survey of Outlier Detection Methodologies, *Artif. Intell. Rev.*, 22(2):85-126, 2004.

[3] Irad Ben-Gal, Outlier detection, In: Maimon O. and Rockach L. (Eds.), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, 2005.

[4] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM Comput. Surv.*, 41(3), Art. 15, 2009.

[5] J. H. Friedman and J. W. Tukey, A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Comput.*, C23(9):881-890, 1974.

[6] Stahel, W. A., Breakdown of covariance estimators, Research report 31, Fachgruppe fur Statistik, E.T.H. Zuurich, 1981.

[7] J.H. Friedman, Exploratory Projection Pursuit, *J AM STAT ASSOC*, 82(1):249-266, 1987.

[8] V. Vemuri and W. Cedeńo, *Multi-Niche Crowding for Multimodal Search. Practical Handbook of Genetic Algorithms: New Frontiers*, Ed. Lance Chambers, vol.2, 1995.

[9] A. Ruiz-Gazen, S. L. Marie-Sainte, and A. Berro, Detecting multivariate outliers using projection pursuit with particle swarm optimization, *Proc. of COMPSTAT2010*, 89-98, 2010.

[10] Knorr, E.M. and Ng, R.T., A unified approach for mining outliers, *Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON)*, Toronto, Canada, 1997.

[11] Knorr, E.M. and Ng, R.T., Finding intensional knowledge of distance-based outliers, *Proc. Int. Conf. on Very Large Data Bases (VLDB)*, Edinburgh, Scotland, 1999.

[12] Angiulli, F. and Pizzuti, C., Fast outlier detection in high dimensional spaces, *Proc. European Conf. on Principles of Knowledge Discovery and Data Mining*, Helsinki, Finland, 2002.

[13] Hautamaki, V., Karkkainen, I., and Franti, P.. Outlier detection using k-nearest neighbour graph, *Proc. IEEE Int. Conf. on Pattern Recognition (ICPR)*, Cambridge, UK, 2004.

[14] A. Sierra, High-order Fisher's discriminant analysis, *Pattern Recognition*, 35(6):1291-1302, 2002.

[15] J. Handl, J. Knowles, Feature subset selection in unsupervised learning via multiobjective optimization, *Int. J. of Computational Intelligence Research*, 3:217-238, 2006.

[16] M. Breaban, H. Luchian, A unifying criterion for unsupervised clustering and feature selection, *Pattern Recognition*, 44(4):854-865, 2011.