View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

The Pareto Principle in Datamining: an Above-Average Fencing Algorithm

K. Macek

This paper formulates a new datamining problem: which subset of input space has the relatively highest output where the minimal size of this subset is given. This can be useful where usual datamining methods fail because of error distribution asymmetry. The paper provides a novel algorithm for this datamining problem, and compares it with clustering of above-average individuals.

Keywords: ART, Pareto principle, insurance risk.

1 Introduction

In some cases, usual methods of supervised learning are not able to provide satisfactory results. This may occur in data with asymmetric distributed error, which is typical in insurance. In order to manage the asymmetry in the sense of the law of large numbers, this paper offers a new algorithm, which constructs a predictor not for points, but for sets. We will show an algorithm for finding sets of units with above-average outputs.

Let X be a set, and let μ, ν be measures over it. The Pareto principle arises if there is a set $P \subset X$ where

$$p(P,X) \equiv \frac{\nu(P)}{\mu(P)} \cdot \frac{\mu(X)}{\nu(X)} \gg 1 \qquad (1)$$

and $r(P) \equiv \mu(P)/\mu(X) \gg 0$. Let μ stand for volume, ν for production, p for productivity, r for proportion. Typically, the Pareto principle is considered as a rule that 20 % of elements "produces" 80 % or more of the output. (This principle was discovered by Vilfredo Pareto while assessing the welfare distribution in the UK at the end of the 19th century. His ideas were systematically described, applied and extended by Max Lorenz[7].) In this case, $r(P) = 0.2$ and $p(P) \geq 4$. Managerial science often works with the Pareto diagram [1]. X is discrete $\{x_1, x_2, \ldots, x_n\}$. The elements are ordered by their production $\nu(x_i)$, and the production is drawn in a chart. In stastitics,the Pareto principle is represented by the continuous Pareto distribution:

$$\Pr(X > x) = \left(\frac{x}{x_m}\right)^{-k} \qquad (2)$$

for all $x \geq x_m$, where x_m is the (necessarily positive) minimum possible value of X, and k is a positive parameter. Pareto distribution has positive skewness

$$\frac{2(1+k)}{k-3}\sqrt{\frac{k-2}{k}},$$

which means that the below-average subset is bigger than the above-average complement. This occurs in many real life situations: the median citizen has a below-average salary, the median driver causes below-average claims, etc.

Let $\xi : X \to \mathbb{R}^n$ be attributes of X. (Attributes can be considered as columns in a data table, i.e. the n-tuple from the i-th row. Mapping ξ can also involve preprocessing. If X is $\mathbb{R}^k$ for some $k \in \mathbb{N}$, the ξ mapping may also be an identity.) The problem of prediction consists in constructing the mapping $\hat{y}$ so that $\int_{i \in X} \left\| \hat{y}(\xi i) - \nu i \right\| d\mu \to 0$.

However, the construction of mapping $\hat{y}$ may be difficult if the Pareto principle arises. A small subset of high productivity (called outliers) corrupts usual assumptions. Usual datamining techniques propose removing the set and working only with the rest. However, in case of the Pareto principle, the small set is very interesting. It is not adequate to speak about outliers, because such data is relevant and obvious. Therefore, we are dealing with more humble result, i.e. with finding the set P defined in (1).

The formulated problem, i.e. finding

$$\arg\max_{P \subset X} p(X,P) \qquad \text{under condition } r(P) > r_0 \qquad (3)$$

is new, and has not been found in the current literature. However, many other topics are related to it. First, clustering methods [12] can be employed. Creating clusters of above-average individuals, the set P will be defined as these clusters.It is important to define the border of the clusters somehow. If these clusters are well found, they can be employed for a more precise approximation [6]. Another approach is to attempt to find a prediction mapping $\hat{y}$ where the set P is afterwards defined at some level of this mapping. RBF neural networks [5] provide an example where the approach of rough sets is employed. Finally, effective P can also be detected also Data Envelopment Analysis [2]. However, none of these methods – as the research in the bibliography shows – has been applied explicitly to the problem of above-average subsets.

2 The Fencing algorithm

The following algorithm is the first attempt to solve this problem (3). It offers the construction of $\hat{P} \subset X$ with above-average production, i.e. with high p. The space $\mathbb{R}^n$ will be

© Czech Technical University Publishing House http://ctn.cvut.cz/ap/

considered as X. The set P is represented by union intervals (from points to hyperboxes) represented by means of complement coding. (Complement coding is a concept applied in ART and ARTMAP neural networks, e.g.[3]. However, the objective of the Fencing algorithm is different. While ART and ARTMAP work iteratively, the Fencing algorithm must often go through the entire training set.) This algorithm works only with finite data sets D, namely with $(\mathbf{x}^i, y_i)$ pairs. Therefore, for all subsets of D, the volume μ is defined as count $\mu(M) = |M \cap \xi(D)|$ and the production as the sum of the production of particular items $\nu(M) = \sum_{i \in \xi^{-1}(M) \cap D} y_i$.

Other ways to construct such intervals may be considered. The following algorithm uses fencing. Fencing is a heuristic approachwhich anticipates that areas of higher average production are located between mutually close points with high production. The algorithm attempts to build a rectangular fence around the area of above-average production, as shown in Fig. 1.

2.1 Measuring and data preprocessing

ξ mapping is necessary. This mapping involves measuring and data preprocessing. The simplest way is to transform binary attributes into real attributes by 0–1 coding. Categorical attributes are transformed into more binary binary attributes. It is very useful to reduce the input vector dimension, e.g. by Principal Components Analysis [11]. Let us define $\mathbf{x}^i = \xi(i)$ and the vector $\mathbf{x}\max = (\mathbf{x}_1^{\max}, \mathbf{x}_2^{\max}, \ldots, \mathbf{x}_n^{\max})$ so $\mathbf{x}_j^{\max} > \mathbf{x}_j^i \ \forall i \in D$. The vector $\mathbf{x}^{\max}$ is used for complementary coding $\mathbf{x}$.

2.2 Data splitting

The data set D is divided randomly into three subsets: base subset B, training subset T and validation subset V. Sets B and T are used for constructing the predictor y, whereby their size will be represented by $|B| \ll |T|$, say $10 \cdot |B| = |T|$. The size of V is chosen with respect to cross validation [8].

2.3 Starting set of intervals

The starting set of intervals is defined as follows $R_O = \mathrm{cc}(xi) | i \in B_O$, whereby $B_O = \left\{ ni \in B | y_i > k \cdot \frac{\nu(B)}{\mu(B)} \right\}$, cc is the complement coding $\mathrm{cc} : \mathbb{R}^n \to \mathbb{R}^{2n}$ and k is a parameter. The definition of BO ensures that $p(B_O, B) > k$.

2.4 Interval expansion

Two intervals $\mathbf{r}^1$, $\mathbf{r}^2$ can be expanded as follows $\mathbf{r}_i^{\mathrm{new}} = \min(\mathbf{r}_i^1, \mathbf{r}_i^2)$, $i = 1, 2, \ldots, 2n$. The construction of $\hat{P}$ consists in iterative expansion of intervals. The process starts with R_O. Two intervals are expanded only if the new interval covers $p(I(\mathbf{r}^{\mathrm{new}}), T) > q_t$, where q_t is a parameter that sinks linearly during the process from p_0 to p_1. In order to expand the compared intervals, a heuristic is used. Two intervals $I(\mathbf{r}^a)$, $I(\mathbf{r}^b)$ are suitable for expansion if $p(I(\mathbf{r}^a))$ and $p(I(\mathbf{r}^b))$ are high, and if $\mathbf{r}^a$ and $\mathbf{r}^b$ are close. Let us define suitability as:

$$v_{a,b} = \frac{p(I(\mathbf{r}^a), T) \cdot p(I(\mathbf{r}^b), T)}{\delta(\mathbf{r}^a, \mathbf{r}^b)}, \tag{4}$$

where δ is a metrics. The first version of the algorithm worked with Hamming distance [6], but other metrics can be also applied. In each step, a pair of intervals is tested for expansion. The pair is selected partly randomly as follows: the pair with highest suitability (probability 0.8), the pair with lowest suitability (0.1), or a random pair (0.1). If a pair is expanded, its suitability is recalculated for all other intervals.

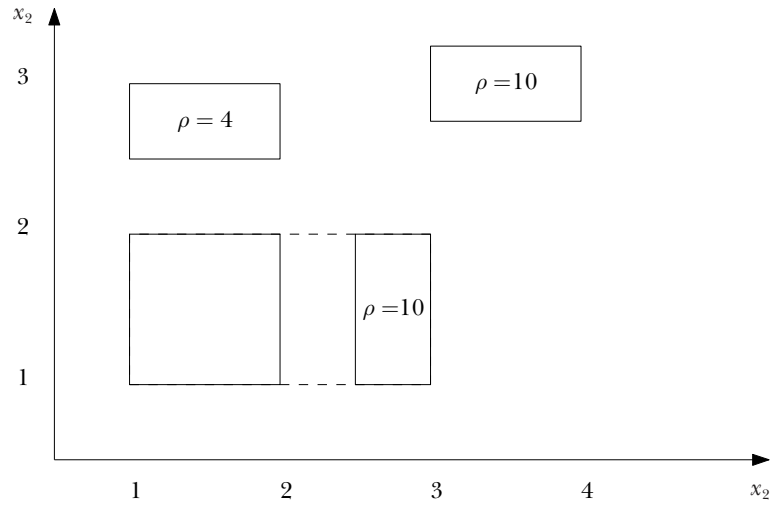

Fig. 1: Fencing: Expanding the square, a new fence (dashed) is recommanded to the rectangle that is close and has high productivity ρ

© Czech Technical University Publishing House

2.5 Termination

The algorithm terminates after all pairs of intervals have been tested and none can be expanded. Afterwards, unexpanded intervals (i.e. points) are deleted. Because intervals may overlap, their conjunction may have lower p than average p of all intervals. Therefore, only intervals with highest p are considered as results so their conjunction has p high enough (e.g. higher than a given threshold).

2.6 Validation

Finally, the results are validated with respect to the validation set $r(P \bigcup V, V)$, and $p(P \bigcup V, V)$ are calculated. Such values can be considered as the quality of the algorithm.

3 Results

The Fencing algorithm has been applied successfully on data on 18 177 insurance claims related to traffic accidents in the Czech Republic in 2003–2005. Categorical and numerical attributes were transformed into binary attributes. There was a total of 135 binary attributes. The considered attributes and their transformation is summarized in Table 1. The Fencing algorithm has been implemented in Matlab as a set of simple scripts. It should be mentioned that this particular data set inspired the author to invent of the Fencing Algorithm, after attempts to build some regression model failed. Generalized Linear Models [4], which are typical in insurance mathematics, and multilayer perceptrons [8] did not provide sufficient results, as shown in Table 2. The data set was split into 10 subsets and one subset was always tested. The logarithm of total costs was taken as an output variable. However, the mean absolute error remains very high (the prediction and reality differ over twentyfold on average!).

First experiments showed that the 135 dimensional space is too sparse and than there are many futher unexpandable intervals. Therefore, the dimension was reduced by selecting 36 attributes describing the region of the claimant, road type, and cause of the accident. After next unsuccessful experiments with $p_1 = 4$ and $p_1 = 2$, it was neccessary to set $p_1 = 1.5$. Then 9 intervals were found. However, the conjunction of them had $p = 1.19$ only. Therefore only 3 best intervals were selected, with $p = 1.48$ and $r = 0.32$.

Table 1: Transformation of observed values into input and output variables

Observed values	Data type	Used as
Accident-related information		
Hour	numerical	24 input binary variables
Day	numerical	7 input binary variables
Month	numerical	12 input binary variables
Year	numerical	6 input binary variables
District	categorical	not used
Municipality size	numerical	6 input binary variables
Cause	categorical	11 input binary variables
Road type	categorical	10 input binary variables
Tariff group	categorical	25 input binary variables
Car make	categorical	not used
Information about causing person		
Age	numerical	8 input binary variables
Sex	categorical	3 input binary variables
District	categorical	not used
Municipality size	numerical	6 input binary variables
Region	categorical	15 input binary variables
Accident at place of abode	binary	1 input binary variables
Claim costs		1 output numerical variable
Paid		numerical
Additional expected		numerical

© Czech Technical University Publishing House

Table 2: Mean absolute error of machine learning for different validation sets

Validation sets #	1	2	3	4	5	6	7	8	9	10
GLM	3.43	3.37	3.27	3.33	3.30	3.22	3.35	3.27	3.43	3.41
MLP	3.35	3.29	3.24	3.30	3.23	3.10	3.27	3.28	3.44	3.34

Table 3: Comparison of the Fencing algorithm with the k-means based approach

Method	r (fixed)	p
means based approach (6 means)	0.32	0.96
means based approach (20 means)	0.32	0.94
Fencing algorithm (raw results)	0.48	1.19
Fencing algorithm (results after best interval selection)	0.32	1.48

3.1 Comparison

The problem (3) formulated here is novel and the Fencing algorithm is the only solution so far. However, for a simple comparison a clustering based method was involved that can be described briefly as follows:

1. **Building above-average clusters from training data**: Best 20 % records were extracted and clustered via the k-means algorithm. For each cluster, the diameter was calculated as the maximum of distance between the center and the record belonging to it.
2. **Finding above-average records in the testing data**: For each record, we test whether there is a cluster whose center is closer to the record than the c multiplied diameter of the cluster. Parameter c is set up so that the level of r is satisfied. So P is defined and r ensured.
3. **Calculation of** p from provided data, p is calculated.

Table 3 shows the results achieved by this method, and compares them with the Fencing Algorithm: the alternative method based on known algorihm provides less narrow results. However, the goal of this paper was not test proposed Fencing Algorithm, but to show that this algorithm is able to solve problem formulated above (3). More experiments with the k-means based approach might provide better results.

4 Discussion and further work

The Fencing algorithm can be modified so the suitability $v_{a,b}$ is calculated in another way. There should be an increase in p in both intervals and a decrease in distance between $r(a)$ and $r(b)$. The randomized selection rule can also be modified.

If a pair of intervals is tested, the whole training set T is gone through. This is probably the Achiles tendon, because the size of T is usually very large. Therefore more detailed examination complexity and the design of more suitable data structures are desirable.

The basic idea of constructing an above-average subset can be evolved in many ways. The subset need not be a union of intervals, but they may be simplexes. The set must not be narrow, it may be fuzzy. Or the subset can be given in an algebraic form and detected by genetic programming or other optimization methods, such as Ant Colony Optimization [10]. The Fencing algorithm will be compared with these other approaches in terms of complexity and effectiveness on more data sets. Systematic examination of relevant preprocessing methods is also desirable. Finally, the algorithm could be modified not for data, but for an estimated probability function, e.g. in form of copulas [9] which are more appropriate for assymetric distributions.

5 Conclusion

The Fencing algorithm is a novel heuristic method for finding a subset of with above-average production. The main idea of the algorithm is to join intervals with high production and small mutual distance. The Fencing Algorithm has been successfully applied to insurance data. Further work has been discussed above.

References

[1] Akpolat, H.: *Six Sigma in Transactional and Service Environments*. . Gower, Burlington, Vt.:,Hasan Akpolat 2004.

[2] Andersen, P., Petersen, N. C.: A Procedure for Ranking Efficient Units in Data Envelopment Analysis. *Manage. Sci.*, Vol. **39** (1993), No. 10, p. 1261–1264.

[3] Dagher, I.: L-p Fuzzy Artmap Neural Network Architecture. *Soft Comput.*, Vol. **10** (2006), No. 8, p. 649–656.

[4] de Jong, P., Heller, G. Z.: *Generalized Linear Models for Insurance Data*. Cambridge Press, 2008.

 © Czech Technical University Publishing House

[5] Jia, Z., Gong, L.: The Project Risk Assessment Based on Rough Sets and Neural Network (rs-rbf). In *Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on Wireless Communications*, 2008.

[6] Kreinovich, V., Yam, Y.: *Why clustering in function approximation?* Theoretical explanation, 2001.

[7] Lorenz, M. O.: Methods of Measuring Concentration and Wealth. *Journal of the American Statistical Association*, Vol. **9** (1905), p. 209–219.

[8] Mitchell, T. M.: *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1997.

[9] Nelsen, R. B.: *An Introduction to Copulas*, volume 139 of *Lecture Notes in Statistics*. New York: Springer-Verlag, 1999.

[10] Ramos, G. N., Hatakeyama, Y., Dong, F., Hirota, K.: Hyperbox Clustering with Ant Colony Optimization (haco) Method and its Application to Medical Risk profile Recognition. *Appl. Soft Comput.*, Vol. **9** (2009), No. 2, p. 632–640.

[11] Theodoridis, S., Koutroumbas, K.: *Pattern Recognition*, Third Edition. Academic Press, February 2006.

[12] Xu, R., Wunsch, D.: *Clustering*. IEEE Press Series on Computational Intelligence. John Wiley & Sons, 2009.

Mgr. Karel Macek
e-mail: karel.macek@fjfi.cvut.cz

Department of Mathematics

Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Trojanova 13
120 00 Praha2, Czech Republic

© Czech Technical University Publishing House http://ctn.cvut.cz/ap/