THE ORIGINAL CLASSIFICATION ALGORITHM FOR THE IMPROVEMENT OF REGRESSION MODELS FOR THE PURPOSE OF TAXATION

E.K. Kornoushenko
Institute of Control Sciences RAS
Moscow, Russian Federation
e-mail: ekorno@mail.ru

A.A. Lobko
Moscow Institute of Physics and Technology
Moscow, Russian Federation
e-mail: alex.lobko@gmail.com

Abstract

Quality of a practical regression appraisal in some cases can be improved by definition and subsequent division of the market sample into two latent classes of "cheap" and "expensive" objects and by constructing models for corresponding classes. Before the calculation of the objects' prices each of them has to be ascribed to one of two classes defined. Several methods of classification were analyzed and original algorithm, named KL, was developed, which possesses a few important advantages in comparison with recognized kNN and C4.5 algorithms. Described approach has proven effective on real data used in mass appraisal. Essentially, low error rate of classification determines high quality and fairness of regression appraisal for the purpose of taxation.

1 Introduction

One of the favorite scientific approaches to the real-estate mass appraisal is the building of the regression models. This method has unquestionable merits but runs into problems when it is used on the developing markets with heterogeneous samples of real-estate property. A developing or weak market is usually characterized by instability and inadequacy of prices. It is common for the sample to contain insufficient information to explain the discrepancy in price level between seemingly similar objects. In such cases the development of single regression model often results in a low quality of appraisal, in particular on the test sample. Clearly the adequacy of the appraisal is of the paramount importance in the discussed field as these results are explicitly used in the calculation of taxation payments.

One of the possible solutions (using the assumption that the training sample is big enough to divide it) is to break down the initial sample into several classes and to build corresponding models within the respective classes. The whole process is clear if the rule for differentiation is based on a geographical position or any other factual information. However, in most cases the best results are derived when two classes are so-called "cheap" and "expensive" objects - and usually the type of the object is

defined only by a complex combination of attribute values and cannot be described by simple differentiation. Therefore at the outset every object with unknown price (from a test sample) has to be attributed to the one of two classes before the price can be calculated using the model for the corresponding class. As the result the overall appraisal quality greatly depends on the characteristics of the classification algorithm. In this paper the original algorithm (it was called KL after first letters of authors' last names) is presented which showed itself to good advantage in the practical mass appraisal in Russian Federation. Conceptually this algorithm is somewhat close both to C4.5 algorithm developed by Ross Quinlan (1993) and to classic kNN (k-nearest neighbor algorithm)(Cover and Hart, 1967). The main features of the algorithm being presented are firstly the analysis of the informational significance of specific values of attributes (C4.5 uses the concept of the informational gain of the attribute on the whole) and secondly the process of the convergent search for the cloud of resembling objects, which reminds about kNN. Moreover, the informational significance of attribute values is recalculated on every step of the algorithm along with the decrease of uncertainty about the correct class. Thus, one of the most significant characteristics of KL is that algorithm takes into account combinations of attribute values in common. The essence of such approach was briefly described in (Kornoushenko, 2008).

2 Algorithm description

The main idea is that not every value of the same attribute contains the same amount of information about the membership of the certain class. Consider wall material for the house. Both cheap and expensive houses can be constructed out of brick and concrete with almost similar probability. On the other hand, in the country wooden house usually means cheapness. Similar mechanism works for continuous descriptors.

Therefore the key notion of the algorithm is the informational significance S of an attribute value. Let x_{ij} represent value number j of discrete attribute i; n_0 and n_1 are initial numbers of training sample objects in two classes respectively; $N(x_{ij}) = (N_0, N_1)$ is the distribution between the classes C_0 and C_1 of that part of training sample, in which attribute i has value x_{ij}. Then

$$S(x_{ij}) = \frac{max\left(\frac{N_0}{n_0}, \frac{N_1}{n_1}\right) - min\left(\frac{N_0}{n_0}, \frac{N_1}{n_1}\right)}{max\left(\frac{N_0}{n_0}, \frac{N_1}{n_1}\right)}.$$

In case of the continuous attribute the notion of d-closeness is used: value x_{ik} is considered d-close to the value x_{ij} if

$$|x_{ik} - x_{ij}| < d.$$

The expression for S stays valid, only N_0 and N_1 mean the number of objects in two classes which have values, that are d-close to x_{ij}. For the simplicity and to have only one variable d for all continuous attributes such descriptors are overdetermined so that they vary within the same limits. The number d is estimated empirically during assessment of the algorithm performance on training sample.

Test sample consists of the objects that have the same pattern of attributes as the training sample but are of unknown class membership. S is calculated for every attribute value of the object being classified. After that the training sample is filtered down so that only objects with either matching or d-close values of the most significant attribute are left. For this selected group of objects significance of values is recalculated. Thus the algorithm provides the ability to assess the importance not only of certain values but also of combinations of attribute values considered in common, which contribute largely to the selection of the class.

These procedures of convergent filtration continue until one of next several conditions is satisfied:

- Selected group contains objects from single class. In this case the process for the test object is finished;
- The choice has not been made though the training set was filtered along all of available descriptors;
- There are some not used attributes left, but during last iteration of filtration not a single similar object was selected.

Last two variants are quite uncommon, but they still have to be dealt with. There are several possible solutions. First of all, such problems with definite classification signal that this object cannot be explicitly ascribed to any of the classes. Then it is perfectly appropriate to use arithmetic mean of two models for the test object in question. Secondly, it is possible to make the choice based on the majority in the last iteration in which there were two classes present in the selected group. Thirdly, the decision can be made on the basis of the most significant attribute - the most probable class for the certain value of the attribute is the result of classification in this case. In fact, these different approaches may change outcome for each single object, but on average they give similar error rate on different test samples.

In (Kotsiantis, 2007) extensive review of classification algorithms is presented and it is stated that basically there are no better or worse algorithms (among good algorithms) - some algorithms are just more suitable for certain tasks. According to (Hyunjoong Kim et al., 2007) in real-estate appraisal decision trees (and in particular C4.5) show best results among whole set of available algorithms.

To compare algorithms it is useful to discuss their disadvantages. kNN is very sensitive to irrelevant features because of the way it works - in contrast, the attentive selection of significant attributes in KL allows to cope with such problem. Moreover kNN is intolerant of noise and can be easily distorted by errors in data. Contrary to kNN, decision trees and KL algorithm are more resistant to noise because their strategies help to avoid overfitting.

On the other hand kNN and KL require only few parameters to tune. Next, decision trees cannot perform well with problems that require diagonal partitioning as the division of the instance space is orthogonal to the axis of one variable and parallel to all other axes. Hence, the resulting regions after partitioning are all hyperrectangles. KL algorithm with its attention to specific values and not attributes on the whole does

not have this problem. Surely, it is not very fast but in a weak market there is no such problem as big volume of training sample. Much more often it is the other way round.

It is yet to be understood what precise features of the sample are needed for KL to outperform other classification algorithms. However several trials both of KL and C4.5 (which is arguably the best algorithm for the real-estate data) on data from official sources (information that was used in research was actually utilized for mass appraisal in Russian Federation) showed that in most cases the error rates are similar. There are samples on which KL tends to perform slightly better than C4.5.

3 Conclusion

Work in the area of mass appraisal for the taxation purposes has led to the realization that definition of latent classes in the market sample can substantially increase prediction power of the composite model in comparison to the single one.

In an attempt to better understand strengths and limitations of different classification methods several of them were investigated and as the result KL algorithm was developed. Though KL is not easy to interpret, its essence is rather transparent. The principles of KL are lucid and are in many ways similar to the reasoning of sensible human trying to solve such a problem intuitively. In summary, high-quality results (comparable to that of C4.5 and in some cases better) of KL algorithm derived on real data aggregated in Russia prove efficiency and functionality of the described approach to classification. Such results allow to conduct fair appraisal of estate property market price, which is the basis for the imposition of socially meaningful taxes.

References

[1] Cover T.M. and Hart P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. Vol. **13(1)**, pp. 21-27.

[2] Hyunjoong Kim, Wei-Yin Loh, Yu-Shan Shih and Probal Chaudhuri (2007). Visualizable and Interpretable Regression Models With Good Prediction Power. *IIE Transactions*. Vol. **39**, pp. 565-579.

[3] Kornoushenko E.K. (2008). Methodology of practical regression analysis. *Journal of Control Sciences*. Vol. **2**, pp. 34-41, (in Russian).

[4] Kotsiantis S.B. (2007). Supervised Machine Training: A Review of Classification Techniques. *Informatica*. Vol. **31(No.3)**, pp. 249-268.

[5] Quinlan Ross (1993). *C4.5: Programs for machine training*. Morgan Kaufmann, San Mateo, CA.