Instance-Based Classifiers for Tax Agent Modelling

Fuchun Luan and Warwick Graco, Australian Taxation Office, and Mark Norrie, YiSi Technologies, Canberra, Australia

---

Tax agents are responsible for assisting taxpayers to submit tax returns for individual, company, and other types of tax returns and to prepare business activity statements for business taxes, including goods and services taxes or GST. Unfortunately, some agents abuse their positions of trust to defraud the tax system. One way they do this is by inflating the business deductions of their clients. The Australian Taxation Office (ATO) is responsible for identifying high-risk tax agents who are engaging in unacceptable practice. The methods described in this paper were aimed at identifying high-risk agents.

In this paper, we report some results from modelling tax agent behavior using a distance-from-the-centroid (DFC) method with assistance from a genetic algorithm (GA). DFC is an example of what are called "instance-based learning methods." These use known high-risk cases, or instances, to see if other cases have practice profiles that are similar to them.

DFC works simply by identifying the center of gravity or centroid of a collection of known high-risk cases and then finds other cases not previously classified that are close in distance to the centroid. GAs are ideal for problems which require optimized solutions (Goldberg, 1989). They have been successfully applied to a great variety of real world problems, including timetabling, job assignment, and travelling salesman problems (Luan and Yao, 1996). In the present study, they are employed to optimize the weights of the attributes which discriminate between known high-risk cases and those whose risk classifications are not known. GAs use Darwinian survival of the fittest to breed offspring (which in this research are new sets of variable weights) that help distinguish between the two categories of cases. This reproduction process continues until an optimized set of weights is found.

The remainder of this paper will report some initial results from using DFC. This is followed by an outline of other instance-based methods that are being investigated by the Analytics Group at the ATO. Other pertinent issues to do with classification modelling are briefly covered, and some of the research into instance-based methods is highlighted.

DFC Method

Subjects

The steps here included:

- 14,913 agents were selected for Income Year 2002. These were active agents who practiced throughout the year.
- 49 known cases of high-risk agents were nominated by ATO compliance staff and were used as a high-risk group in the research. These agents were mainly those who manage the tax affairs of individual taxpayers. Only a few agents who deal with company, partnership, and trust clients were nominated in this collection of high-risk agents.

Data

The data used were extracted from the ATO enterprise data warehouse for Income Year 2002. The research focused on examining the characteristics of tax agents via their aggregated clients' tax return data. Data on 256 variables (also called "attributes" or "features") were used in the research. The variables included descriptive and summary statistics of tax agent practice, such as total number of clients serviced and average deductions claimed for rental property.

Feature Extraction

The 256 variables were far too high a number for the DFC modelling that was carried out. It is very difficult to develop effective models when the data have high numbers of variables. Steps were taken to identify variables which discriminated the high-risk tax agent group from other agents in the population. A comparison was made between the mean values of the variables for the high-risk group with those of the remaining agents. It was found that up to 16 variables distinguished between the two groups (see Figure 1). These discriminating features cannot be listed for confidentiality reasons. However, they covered such issues as high-risk tax agents inflating claims for work-related expenses and deductions for rental properties compared to other agents.

Profiling and Modelling

The DFC modelling techniques rank ordered all tax agents based on the distance their profiles were from the centroid of the profiles of the group of high-risk agents (see Figure 2). The discriminatory variables used to determine the distance scores were weighted based on the degree they maximized the pickup rate of the high-risk agents in the 500 highest ranked profiles. This was to ensure

Figure 1. Tax Agent Profile Benchmarks

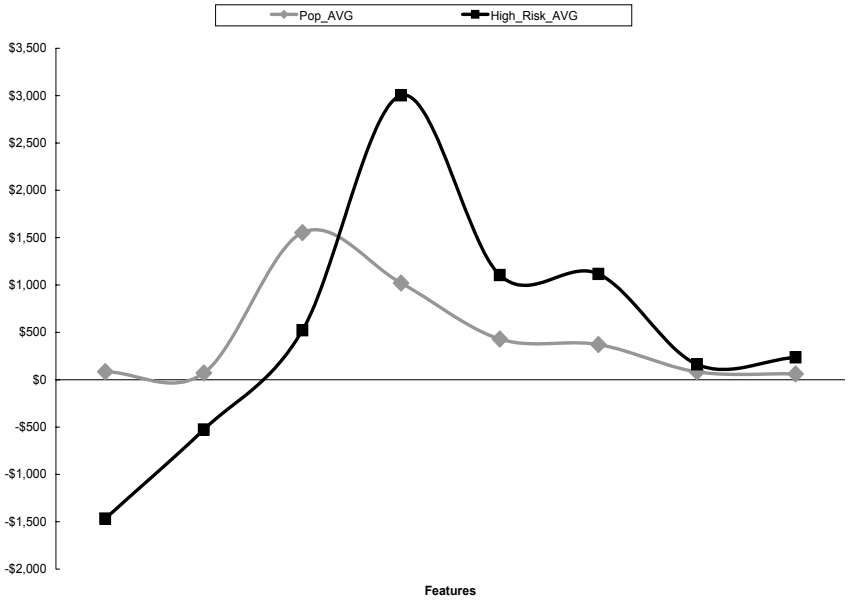
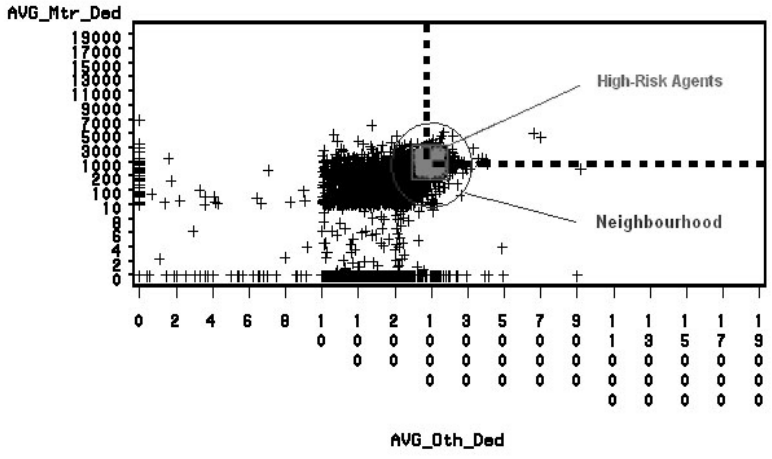


Figure 2. The Square represents the locations of 49 high-risk agents. The size of the entire population is 14,913.



that the top group of high-risk agents was clearly seen in the data because they were the group of most interest to the ATO.

Procedures for Calculating DFC

These included:

- The discriminatory variable mean value was calculated to give a score for each agent. The entire population of 14,913 tax agents were profiled and ranked based on each individual agent's score, which is calculated based on his or her location in relation to the center of the known high-risk agent cluster (see Figure 1).
- GA was employed to optimize the weights applied to the various discriminatory variables. The aim of using a GA is to weight higher those variables which are more discriminatory.
- All agents are scored using the weighted discriminatory variables.

The scoring formulae used in the DFC calculations were:

$$S_j = \sum_{i=1}^{n} \left(Wi \times \frac{(Fij - \overline{Fi})}{\overline{Fi}} \right) **2 \qquad (1)$$

where *i* is the *i-th* selected variable (column), *j* is the *j-th* tax agent (row), and W_i is the weight, and $\overline{Fi}$ is the mean value of *i-th* feature for the high-risk group.

The closer the tax agent profiles were to the mean profile of the high-risk group for the weighted discriminatory variables, the lower their DFC scores. The lower the score, the higher the risk the tax agent was practicing in a manner that was unacceptable. All 14,913 profiles were scored and ranked in this manner.

Results

The top 500 agents selected using the DFC method included 40 out of 49 high-risk agents. This gave an 82-percent pickup rate.

Discussion

The results showed that:

- Only a small number of variables (in our case 16) out of a possible 256 were found to discriminate between 49 high-risk tax agents and the remaining population of 14,864 tax agents.
- The discriminatory variable scores of the 49 high-risk tax agents formed a tight cluster with relative low spread or variance (see Figure 2).
- The difference in the mean values of the discriminatory variables between the high-risk cluster and that of the general population of tax agents was more than double.
- The DFC has the advantage that it can rank order the entire tax agent population.

One issue which was not explored further in the research was the outlying cases that had high scores for the discriminatory variables (those that would be located to the top right-hand quadrant in the top graph of Figure 2), thus suggesting that they could be abusing the tax system. A formula for identifying agents in this quadrant is:

$$S_j = \sum_{i=1}^{n} \{SFiWi \times \frac{(Fij - \overline{Fi})}{\overline{Fi}}\} **1 \qquad (2)$$

where *i* is the *i-th* selected variable (column), *j* is the *j-th* tax agent (row), and Wi is the weight, $\overline{Fi}$ is the mean value of *i-th* feature for the high-risk group, and SFi is sign flag.

The cases in this quadrant were not reviewed by compliance staff. However, it has been found at the ATO that cases with outlying scores often have understandable reasons for their unusual profiles, such as they service particular types of clientele. Cases which are more likely to be of concern to the ATO are boundary ones. These cases are on the border of unacceptable practice and manage their affairs so that they are less likely to be detected.

Other Research

The DFC is one type of instance-based learning. There are others that have been researched for identifying noncompliance. One is the traditional k near-

est neighbor (KNN) method and the other a modification of this called a radial KNN (RKNN).

KNN finds a "k" number of cases specified by the user that are closest to a known high-risk case. For example, the user may want to find the five closest neighbors (ie k=5) to each known instance. If there were 10 known high-risk cases, this would provide a total of 50 nearest neighbors (i.e., 10 known cases * 5 nearest neighbors).

This method has a number of drawbacks, including, firstly, there can be multiple instances where the same case is identified as a nearest neighbor to two or more known cases. Secondly, a case may be the nearest neighbor to a known high-risk case but still be a considerable distance from it. Thirdly, this algorithm does not include categorical variables in its calculations. For example, the type of industry where a taxpayer operates could be a discriminator and can assist to ensure cases are correctly classified. Industry codes can be used in the RKNN calculation.

The RKNN[1] overcomes all three weaknesses of the KNN. It ensures that each nearest neighbor identified is not duplicated with other known high-risk cases. It specifies a circle around which a case variable must be distant from a known case as shown in Figure 3. Cases located inside the circle are classified as nearest neighbors. Those outside the circle are not as shown in Figure 3. This algorithm also includes categorical variables in the calculation of the nearest neighbor. The RKNN is currently being evaluated. We are also investigating if RKNN outperforms KNN and DFC.

The obvious question that could be raised is why these different algorithms were developed and tested by the ATO. The simple answer is that, when we started using instance-based learning methods, there were no commercial off-the-shelf methods readily available and so, the DFC was developed initially as a stop-gap measure. This was followed by the traditional KNN when access to a commercial algorithm was gained. This algorithm was found to have the deficiencies stated above, and this led to the development of the RKNN.

Other Issues

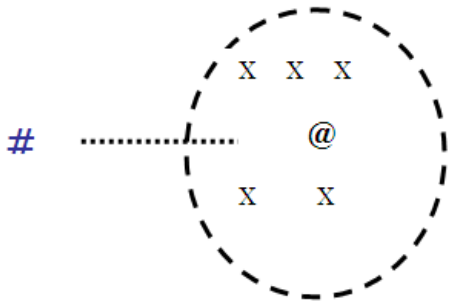
Instance-based methods have a number of advantages including that they are simple and intuitive to use and understand and they are learned quickly and provide good accuracy for a variety of real-world classification tasks. However, they do have weaknesses, including that they can require large storage because they store the training data, they can be computationally intensive because all training instances must be searched in order to classify cases, and they are susceptible to what is called "the curse of dimensionality." This is where there are too many variables in the data. They are also affected by error or noise in the data.

The most attractive characteristic of this form of learning is that the classifier can be developed quickly using a small number of known high-risk cases. This is in contrast to other types of classifiers that usually require medium-to-large numbers of classified cases to perform well.

From an administrative point of view, there are other challenges with using these and other types of classifiers. One is that users may experience difficulties understanding why cases were classified as potentially high-risk. While instance-based methods may be transparent in the way they operate, they are not always transparent with the reasons why cases are identified as potential risks. One lesson learned at the ATO is that it is very important to explain why cases are considered to be potential high-risks to those who do audits and investigations. Many classifiers use general models that indicate which cases are at risk based on practice statistics, such as profit to income and costs of goods and services to turnover. These statistics do not always make sense unless they are related to industry norms, such as which industry each high-risk case operates.

What has also been learned at the ATO is that a good case-selection tool is required to convert the results of general models into specific audit and investigatory issues that compliance staff can take forward in their compliance work. If this tool is not available, compliance staff can struggle to understand the models. From this perspective, a case-selection tool is integral to the models in that the two go together like a hand in a glove.

Figure 3. This shows that only neighbors inside a circle are considered with RKNN and a Case such as # which is outside the circle is ignored.



Another lesson learned at the ATO with modelling is that it is better to develop single-issue models, such as for shareholder loans to company directors, capital gains, work-related expenses, and rental income. Single-issue models are easy to develop, are easy for compliance staff to understand, and are easy to audit/investigate issues identified by the models.

One misconception we encountered in the ATO is the belief that the models are only suitable for high-volume, simple tax issues and that they are unsuitable for complex and difficult tax matters such as found with large multinationals. This is a misunderstanding of the power of models. Complex tax issues can be broken down into simpler, single issues and a model developed for each one. Furthermore, it has been found at the ATO that, while single-issue models can appear in some cases to be weak or trivial in that they lack discriminatory power, when combined, they can be powerful classifiers. That is, there is strength in numbers with classification models.

It has also been found that there can be overflows or spillovers with the model results. These are additional benefits that the models were not designed to deliver. One type of overflow is where the models point to other issues besides those the model was designed to provide. For example, a model might have been developed to identify business clients who have serious debt problems and will struggle to repay money owing to the ATO. These models can also indicate that these clients may not forward the income tax they collect from their employees each pay period to the ATO.

Another type of overflow is one where tax agents who normally manage large and medium business clients are identified to have potential compliance problems with their microbusiness clients. This suggests that, if they are having compliance problems with this type of client, they should be checked to see if they are having problems with their other types of business clients.

There have been other developments with instance-based classifiers. They include:

- The use of unclassified cases to improve KNN performance (Driessens et al., 2006). The researchers used another classifier to preclassify a selected number of unknown cases. These newly classified cases were then combined with the known classified cases to develop the KNN classifier. It was reported that this improved the performance of the classifier.
- The development of algorithms that overcome storage and performance problems of KNN (Ritter et al., 1975; Wilson and Martinez, 2000).
- The use of performance bias methods and preset bias methods[2] for feature selection for KNN. Performance bias methods, which are

also called "wrappers," find a set of feature weights through an iterative procedure that uses the classifier's feedback to improve the weights. Preset bias methods, also called "filters," use a predetermined function that measures the information content of each feature, and features are selected based on their information yield. The higher the yield, the better the feature.

- The application of bucket or grid methods (Yianilos, 1993) that divide the distribution of unknown cases into identical cells. The cells are examined for presence of neighbors in order of increasing distance from a known case or instance. The search terminates when the distance from the known case to the cell exceeds the distance to the closest unknown case already visited.
- The generation of what are called k-d trees (Friedman et al., 1977). These are binary trees that divide unknown cases into multidimensional rectangles using the feature scores until the number of cases in each rectangle is below a given threshold. This approach assists to speed up KNN search.

Conclusion

Instance-based methods are simple and easy to use and can provide quick results with classification of cases. They do however have a number of technical and administrative challenges. It is recommended that to obtain the best results from these methods that they be restricted to issues that are relatively simple and straightforward, that care be taken to identify and use the features that discriminate between high-risk and low-risk cases, and that tight matching requirements be imposed between known high-risk cases and their nearest neighbors. It was also recommended that single issue models be produced as these are easier to develop and easier to implement and that boundary rather than outlying cases should be detected as these are more likely to be noncompliant.

Endnotes

[1] This algorithm was developed by Tatiana Semenova from the Analytics Group at the ATO.

[2] This was reported in a lecture on Nearest Neighbors by Professor Riccardo Gutierrez-Osuna at Texas A&M University. See http://research.cs.tamu.edu/prism/lectures.htm

References

Driessens, K; Reutermann, P; Pfahringer, B; and Leschi, C. (2006), Using Weighted Nearest Neighbor to Benefit from Unlabeled Data, in W.K. Ng; M. Kitsuregawa; J. Li; and K. Chang (editors), *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin, pp. 60-69.

Friedman, J; Bentley, J; and Finkel (1977), An Algorithm For Finding the Best Matches in Logarithmic Expected Time, *ACM Transactions on Mathematical Software*, Number 3, pp. 209-226.

Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimisation, and Machine Learning*, Addison-Wesley, Reading, MA.

Luan, F.C. and Xin, Y. (1996), Lecture Room Assignment with Genetic Algorithm, in R. Stocker; H. Jelinek, B. Durnota; and T. Bossomaier, T (editors), *Complex Systems*, Number 96, pp. 149-160 (see http://journal-ci.csse.monash.edu.au/ci/vol03/luanyao/)

Ritter, G.L; Woodruff, H.B; Lowry, S.R; and Isenhour, T.L. (1975), An Algorithm for a Selective Nearest Neighbor Decision Rule, *IEEE Transactions in Information Theory,* Number 21, pp. 665-669.

Wilson, D.R. and Martinez, T.R. (2000), An Integrated Instance-Based Learning Algorithm, *Computational Intelligence*, Number 16, pp. 1-28.

Yianilos, P.N. (1993), Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces, *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 311-321.