Identifying and Characterising Anomalies in Transaction Data

Koen Smets Jilles Vreeken

Universiteit Antwerpen, Belgium

Abstract

In many situations there exists an abundance of positive examples, but only a handful of negatives. In this paper we show how in binary or transaction data such rare cases can be identified and characterised.
Our approach uses the Minimum Description Length principle to decide whether an instance is drawn from the training distribution or not. By using frequent itemsets to construct this compressor, we can easily and thoroughly characterise the decisions, and explain what changes in an example would lead to a different verdict. Furthermore, we give a technique through which, given only a few negative examples, the decision landscape and optimal boundary can be predicted—making the approach parameter-free. Experimentation on benchmark and real data shows our method provides very high classification accuracy, thorough and insightful characterisation of decisions, predicts the decision landscape reliably, and can pinpoint observation errors. Moreover, a case study on real MCADD data shows we provide an interpretable approach with state-of-the-art performance for screening newborn babies for rare diseases.

1 One-Class Classification by Compression

By running KRIMP, an itemset-based compressor [4] on training database D, a transaction database consisting of positive examples, we obtain an approximation of the MDL-optimal compressor for D which we employ for one-class classification, or outlier detection. If the encoded length L of a transaction is larger than a given threshold value, we decide it is an outlier. Formally, this means that given a decision threshold θ, a code table CT for database D, both over a set of items $\mathcal{I}$, for an unseen transaction t also over $\mathcal{I}$, we decide that t belongs to the distribution of D iff $L(t \mid CT) \leq \theta$.

Cantelli's inequality gives us a well-founded way to determine a good value for the threshold θ; instead of having to choose a pre-defined *amount* of false-negatives, we can let the user choose a confidence level instead. That is, an upper bound for the false-negative *rate* (FNR). For instance, a confidence level of 10%, corresponds setting θ at 3 standard deviations to the right of the average. This means the user has less than 10% chance of observing a future positive transaction that lies further than the decision threshold.

1.1 Characterising Decisions

The main advantag of using a pattern-based compressor like KRIMP, is that we can *characterise* decisions. As an example, suppose a transaction t is classified as an outlier. That is, $L(t \mid CT) > \theta$. To inspect this decision, we can look at the itemsets by which the transaction was covered; this gives us information whether the outlier shows patterns characteristic for the positive class. That is, the more t resembles the patterns of the positive class, the more it will be covered by long itemsets and less by singletons. On the other hand, patterns that are highly characteristic for D that are *missing* from the transaction cover are equally informative; they pinpoint where t is essentially different from the positive class.

Since code tables on average contain up to a few hundred of elements, this analysis can easily be done by hand. In addition, we can naturally rank these patterns on encoded size, to show the user what most characteristic, or frequently used, patterns are missing or present. As such, decisions can easily be inspected.

This is an extended abstract of our work published as K. Smets and J. Vreeken, The odd one out: Identifying and characterising anomalies. In *Proceedings of the SIAM International Conference on Data Mining (SDM'11)* [3].

Koen Smets and Jilles Vreeken are respectively supported by Ph.D. and Postdoctoral Fellowships of the Research Foundation - Flanders (FWO) and thank i-ICT of Antwerp University Hospital (UZA) for providing MCADD data and expertise.

1.2 Estimating the Decision Landscape

For many situations it is not unrealistic to assume that, while not abundant, *some* example outliers are available besides the training data (e.g. less than 10). Even if these examples are not fully representative for the whole negative class $\mathcal{D}_n$, we can use them to make a more informed choice for the threshold parameter. To this end, we propose to generate artificial outliers, based on the given negatives, to estimate the number of bits our positive-class compressor will require to encode future samples from $\mathcal{D}_n$; given this estimated distribution of encoded lengths, and the encoded lengths for the training data, we can set the decision threshold θ to maximise expected accuracy—or to inspect whether it is likely we will see good classification scores.

1.3 Measuring Decision Certainty

It is also useful to show how a transaction t needs to be modified in order to change the classification verdict. Or, to show what items are most important with regard to the decision of t. We look therefore at the certainty of a decision by considering the encoded lengths of altered transactions. The rationale is that the more we need to change t to let its encoded length reach below the decision threshold, the more likely it is this example is indeed an outlier. Alternatively, for a sample with an observation error, a small change may be enough to allow for a correct decision.

2 Case study: MCADD

To demonstrate the usability of our approach in a real problem setting, we perform a case study on real MCADD data, obtained from the Antwerp University Hospital. Medium-Chain Acyl-coenzyme A Dehydrogenase Deficiency (MCADD) [1] is a deficiency newborn babies are screened for during a Guthrie test on a heel prick blood sample. This recessive metabolic disease affects about one in 10 000 people while around one in 65 is a carrier of the responsible mutated gene. If left undiagnosed, this rare disease is fatal in 20 to 25% of the cases and many survivors are left with severe brain damage after a severe crisis.
In our study, the dataset contains controls versus MCADD, with respectively 32 916 negatives and only 8 positives. Repeated experiments using 10-fold cross-validation show that all 8 positive cases are ranked among the top-15 largest encoded transactions. Besides, we notice that the obtained performance indicators (100% sensitivity, 99.9% specificity and a positive predictive value of 53.3%) correspond with the state-of-the-art results[1, 2] on this problem. Moreover, analysing the positive cases by manually inspecting the patterns in the code table and covers, reveals that particular combinations of values for acylcarnitines C2, C8 and C10 together with particular following ratios $\frac{C8}{C2}$, $\frac{C8}{C10}$ and $\frac{C8}{C12}$ were grouped together in the covers of the positive cases. Exactly these combination of variables are commonly used in diagnostic criteria by experts and were also discovered in previous in-depth studies [1, 2]. The largest negative samples stand out as a rare combination of other acetylcarnitine values. Although these samples are not MCADD cases, they are very different from the general population and are therefore outliers by definition.

References

[1] C. Baumgartner, C. Böhm, and D. Baumgartner. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *Journal of Biomedical Informatics*, 38:89–98, 2005.

[2] S. Ho, Z. Lukacs, G. F. Hoffmann, M. Lindner, and T. Wetter. Feature construction can improve diagnostic criteria for high-dimensional metabolic data in newborn screening for medium-chain acyl-coa dehydrogenase deficiency. *Clinical Chemistry*, 53(7):1330–1337, 2007.

[3] K. Smets and J. Vreeken. The odd one out: Identifying and characterising anomalies. In *Proceedings of the SIAM International Conference on Data Mining (SDM'11)*.

[4] J. Vreeken, M. van Leeuwen, and A. Siebes. KRIMP: mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1):169–214, 2011.