Performance Analysis and Prediction for Data Mining Systems

Dr. Jorge E. Tierno

BAE SYSTEMS Advanced Information Technologies
6 New England Executive Parkway
Burlington MA 02148
jorge.tierno@baesystems.com

Abstract

We establish theoretical limits on the performance of certain data mining algorithms based only on the properties of the data sets being considered. We demonstrate the use of the bounds with an example based on data generated by an artificial world simulator. We point to extensions of this work and to connections with other fields.

Introduction

Data mining techniques can discover and extract hidden patterns about terrorist activities buried in large data stores, or so it is conjectured. However, given the financial and social costs of collecting and processing such data, it is incumbent upon those responsible for homeland security to evaluate the potential benefit of such data mining systems, namely, assess the ability to detect rare threat events and not to produce a large number of false alarms. We describe techniques from information theory that allow one to estimate performance before actually building the data mining system. These methods can also help to identify what kinds of information are most useful for detecting specified threat patterns.

Technical Approach

Objective

We would like to establish a basis for evaluation of the performance of data mining algorithms. To do so, we set out to determine how much information is provided in the evidence relative to the question we are trying to answer (e.g. classification of a "case" as threat or non-threat). Specifically, given a sequence of events describing a case, we seek to measure how much additional information is needed, in order to always correctly classify it.

The key to solving this problem is to cast it into a framework that allows us to formally define and quantify the information contained in a dataset, as it relates to the detection task at hand. We will use the formalism of Information Theory to define and quantify information

We will approach our objective in three stages. First we will develop an information theoretic metric for data sets. Second we will develop an information theoretic metric for classifiers. And finally we will establish a connection between the two. The approach presented here builds on the work in White et al. 2004.

Brief Review of Information Theory

Given a discrete random variable X, with probability mass function p(x), we define the entropy of X, measured in bits as:

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \quad (1)$$

The entropy can be interpreted as a lower bound on the average number of bits necessary to encode realizations of the random variable X.

Given two random variables X and Y, the conditional entropy of X given Y is defined as

$$\begin{aligned} H(X/Y=y) &= \sum_{x \in X} p(x/y) \log_2 \frac{1}{p(x/y)} \\ H(X/Y) &= \sum_{y \in Y} p(y) H(X/Y=y) \end{aligned} \quad (2)$$

The conditional entropy can be interpreted as a lower bound on the number of bits necessary to encode realizations of X if Y is known. This leads to the definition of mutual information, that captures how much of the variable X is known, by knowing only the variable Y and the joint distribution p(X,Y). The mutual information is defined as:

$$I(X;Y) = H(X) - H(X/Y) \quad (3)$$

The mutual information is the difference between the information in X and the information left in X once Y is known.
Applying Bayes' rule provides the following alternative expressions for the mutual information

$$\begin{aligned} I(X;Y) &= \sum_{y\in Y}\sum_{x\in X} p(x/y)p(y)\log_2\frac{p(x/y)}{p(x)} \\ &= \sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{y\in Y}\sum_{x\in X} p(y/x)p(x)\log_2\frac{p(y/x)}{p(y)} \end{aligned} \quad (4)$$

The relative mutual information is defined as

$$I_r(X;Y) = \frac{I(X;Y)}{H(X)} \quad (5)$$

which measures the fraction of the information in X captured by Y.

Characterization of Classifiers using Relative Mutual Information

An important special case occurs when X is a binary variable and Y is the output of classifier trying to detect the value of X. In this case the conditional probabilities in the last row of Equation (4) are the true positive (Tp), false positive (Fp), true negative (Tn), and false negative (Fn) rates of the detector. Besides depending on the characteristics of the detector, mutual information also depends on the probability distribution of X.
For a given prior p(X), we can compute all the Tp, Fn pairs that correspond to a given relative mutual information. These values correspond to the ROC curve of a classifier with the given Relative Mutual Information. Figure 1 shows an example of these curves for a source with signal to noise ratio (SNR=p(X=true)/p(X=false)) of 1/9.

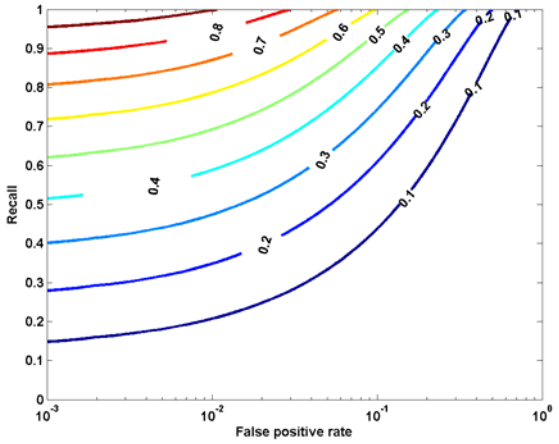

Figure 1: ROC curves as a function of Relative Mutual Information for a classifier with input probability distribution p(X=F)=.9

A common scalar figure of merit for a classifier is the area under the ROC curve (AUC). For a given Input source SNR, we can plot the AUC as a function of relative mutual information.

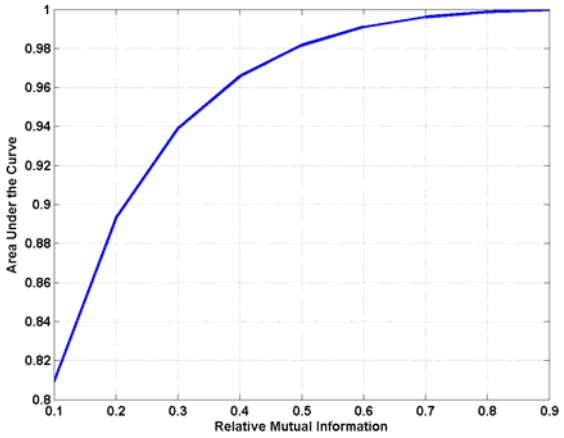

Figure 2: Area under the curve as a function of relative mutual information, for a binary classifier with input SNR=1/9.

As can be seen from Figure 2, these two scalar characterizations of the classifier are in one-to-one correspondence for a specified SNR.

Characterization of Classification Features using Relative Mutual Information

Frequently, we wish to detect a hidden characteristic of a n entity, by relying on some of its observable aspects or features. In this case we will have three random variables to contend with. Variable X encodes the "ground truth" of the entity as to whether it satisfies the hidden characteristic. Variable Y encodes the noisy observation of whether the source contains the feature selected for characterization. Variable Z is the output of our classifier, i.e., our estimation of whether the source verifies the hidden characteristic.
The data processing inequality states that since X →Y→Z we will have (Cover and Thomas 1991)

$$I(X;Z) <= I(X;Y) \quad (6)$$

And thus the classifier based on Y can't provide any more information than Y. The relative mutual information of the feature Y with respect to the input source is thus an upper bound on the performance of the classifier.

Example

As an example we consider relational evidence in an artificial world simulation. In this artificial world groups of people are busy at work carrying out various activities called exploitations. Some of these exploitations are part of the normal productive activity of society. Productive exploitations carried out by legitimate (non-threat) groups are denoted PNT. Productivity exploitations carried out by non-legitimate groups are denoted PT. Finally criminal exploitations are denoted V (and can only be carried out by threat groups). With each of these types of exploitations there is an associated exploitation pattern: the set of activities carried out by the group to consummate the exploitation. Part of these pattern is the set of assets required to consummate them. These sets are called modes. We identify legitimate productivity modes and unlawful vulnerability modes. We will analyze the information content of the different components of the exploitation pattern. Every exploitation is either a V-type, a PT-type, or a PNT-type. We are trying to detect the V-type exploitations. Each of these exploitations is "encoded" in the evidence by a sequence of events, such as team communications (FTC), visit to targets, acquisition of resources and application of assets. We extract one or more features of that pattern to determine how much information it contains (with respect to the exploitation type). For example we will consider the delay between team communication events and team visit events, the number of team communication events, or the type of team communication event used. (See Figure 3)

For each of these pattern components we compute the (relative) mutual information as well as the tell strength as defined in Equation (7). The numerical results are shown in Table 1.

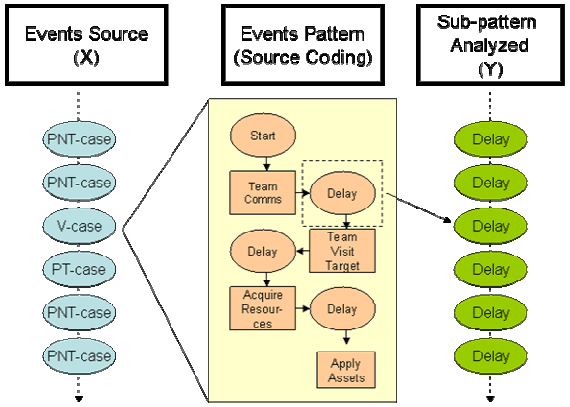

Figure 3: Definition of the variables X and Y used in the analysis of information contained in the Event pattern components.

To compute the relative mutual information with respect to the application of pairs or triples of assets we assume that the productivity and vulnerability modes exploited are picked at random with uniform probability among all the declared modes. If this is not the case the numbers reported can degrade significantly.

For distinguishing between V and not-V we compare mutual information to tell-strength, a metric of the difference between two probability distributions f and g defined as:

$$TS = \frac{1}{2}\sum |f(x) - g(x)| \qquad (7)$$

Sub-pattern Component		Relative Mutual Information		Tell Strength
		V, PT,PNT	V, not-V	
1	1 Delay between FTC events	0.071655	0.077672	0.258966
2	2 Delays between FTC events	0.141408		
3	3 Delays between FTC events	0.208852		
4	1 Delay between 2-ways in FTC	0.078632	0.112209	0.383471
5	Choice of communication type	0.216642	0.242149	0.571109
5a	4 Choices of Communication type	0.596593		
6	Number of cycles in an FTC Event	0.034999	0.029379	0.193191
7	Team Visit Target Event	0.017829	0.024477	0.132577
8	Two asset applications observed		0.396797	0.734418
9	Three asset applications observed		0.844339	0.974580
10	1+4+5+7	0.351227	0.390343	
11	1+4+5+7+8		0.6388	
12	5+9		0.863613	

Table 1: Relative Mutual Information for single and combined features of the pattern

Quantifying Data Association

The results in Table 1 are based on correctly associating all pieces of evidence with their unique cause. This is represented graphically in Figure 4. We will denote the corresponding random variable $\hat{Y}$.

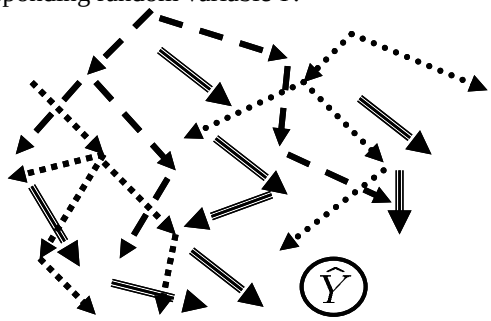

Figure 4: What we need: evidence connected to their exploitation. Line types denote separate cases.

In practice, the datasets will not include case tables for the evidence. This situation is represented in Figure 5. We denote the random variable that commingles the data corresponding to all the cases with the symbol *Y*. To quantify the information loss due to the commingling of the data we introduce an additional random variable *C*, which encodes the association of each piece of evidence with its corresponding case.

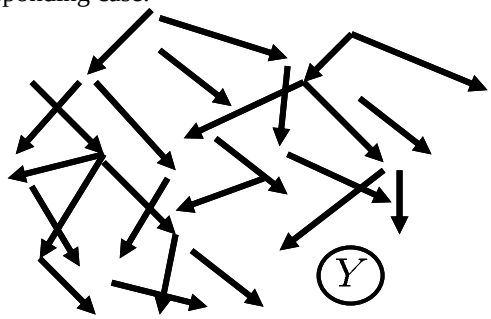

Figure 5: What we have: all evidence commingled. Cases are not labeled.

Applying the transitive property of mutual information we can connect the information content of $\hat{Y}$ and *Y,* as shown in Equation (7).

$$\begin{aligned} I(X;\hat{Y}) &= I(X;Y,C) \\ &= I(X;Y) + I(X;C/Y) \end{aligned} \quad (7)$$

The second term on the right-hand side in Equation (7) is the information lost

$$I(X;C/Y) = H(C/Y) - H(C/Y,X) \quad (8)$$

Connections to System Identification

The problem of bounding the performance of a classifier is akin to the problem of System Identification. In its discrete event form, the most common application of system identification is in the field of Hidden Markov Models. The problem is to determine which model can best predict the output of a target or true system. Information theoretic metrics are used to measure the distance between the target system and the proposed model. Given the target model and the set of models from which to choose a bound on the goodness of fit can be obtained using arguments similar to those in this paper. (See, for example, Ljung 87.)

Conclusions and Challenges

Our research indicates that Information theory can be used to develop consistent metrics for data mining algorithm performance and for the information content of evidence data-bases. These metrics can be used to determine how well a data mining algorithm can possibly perform even before the algorithm is constructed. To further our theoretical developments and to develop practical applications further progress is required.

This analysis assumes that the components of the evidence associated with an event of interest have been correctly identified. We are currently investigating how to measure the effect of clutter and corruption.

Exact computation of mutual information has combinatoric complexity. When computing the relative mutual information due to a set of features, the computation time grows with the product of the number of possible values of each feature. This fact makes exact computation impractical for more than a few features with low cardinality. Approximate computation of Mutual Information is an active field of research (e.g., in bioinformatics), and several new methods have been proposed. These methods need to be extended and applied to the problem of measuring the information on large relational databases.

References

Cover, T., Thomas, J., *Elements of Information Theory*, John Wiley and Sons Inc, NY 1991

Ljung, L. 1987. *System Identification Theory for the User* Englewood Cliffs, New Jersey: Prentice-Hall.

White, J.V., Steingold, S., Fournelle, C.G., 2004. Performance Metrics for Group-Detection Algorithms. In *Proceedings of INTERFACE 2004: Computational Biology and Bioinformatics,* Baltimore, MD.