Chance Discovery

H.M. Hubey

Montclair State University hubeyh@mail.montclair.edu

1 Introduction

Chance discovery is an event with significant impact on human decision making a situation that has the potential for great gain as well as loss. The discovery of a "chance is to become aware of and to explain the significance of a chance, especially if the chance is rare and its significance has been unnoticed" [12]. Conceived this way, it is obvious that probabilistic methods must be combined with decision-making. It can even be interpreted in terms of the old Platonic problem of appearance vs reality, a topic that is dear to philosophers even today. But it is also the problem behind that of cognition/perception in pscyhology, and it is the same problem that occurs in the actual practice of science, and in the discussions of practice of science, e.g. epistemology. In other words, even the discussions in Philosophy of Science revolve around the problem of whether science is akin to black magic or is an efficient and competent extension of the everyday cognition and learning methods that humans employ [5].

In similar ways, one may even consider this problem as that of creativity, albeit not the ones that conventional wisdom calls creativity e.g. art, poetry etc. Scientific creativity is real creativity, and so is the cognition of the importance of events that politicians are faced with, and technological advances that CEOs must contend with in making medium term and long term decisions. Therefore it is expected that a thorough discussion must wend its way through mathematical methods of reasoning, logic, epistemology, cognition, and learning theory.

2 Knowledge

There are a set of related problems in the fields of datamining, knowledge discovery, and pattern recognition. One of them is that we don't know how many neurons should be in the hidden layer or the output layer of Artificial Neural

H.M. Hubey: *Chance Discovery*, Studies in Computational Intelligence (SCI) **30**, 49–68 (2006)
www.springerlink.com

Networks (ANNs) so that for clustering as a preliminary method to finding patterns we must use heuristic methods to determine how many clusters the ANN should recognize. This is just another view of the problem in datamining of knowing how many patterns there are in the data and how we would go about discerning these patterns. There is a related problem in k-nearest-neighbors clustering in which we need an appropriate data structure to be able to efficiently find the neighbors of a given input vector. Indeed, before the k- neighbors method can be used to classify an input vector we need to be able to cluster the training input vectors and an ANN might have been used for this process. The problem of knowing how many patterns (categories or classes/clusters) there are is an overriding concern in datamining, and in unsupervised artificial neural network training. Clustering, datamining, or finding patterns, is then, compression of information, which we call knowledge. And since mathematics is the study of, efficient representation, and compression of patterns, obviously, without mathematics there is no science.

Datamining is based on what was called pattern recognition. One way of classifying the components of pattern recognition is via (i) classification and (ii) estimation. Typically classification is used to create a set of discrete, finite classes, whereas estimation is taken to be an approximation of some desired numerical value based on an observation. The boundaries are not very crisp since estimation consisting of a large number of integer values may just as easily be thought of as categorization or classification. This is especially true if the measured quantities (input data) do not consist of interval or ratio-scaled values. Typically a broad-brush classification of the procedures that consist at least of parts of datamining can be strung along a continuum as shown in Fig. 1.

It may be said that the goal of datamining is to produce domain-knowledge for fields in which there are no models of the type one finds in the ultimate example of a science; physics and its derivatives. An informal listing of the classes of data mining procedures would include, Classification/Segmentation/Clustering, Forecasting/Prediction, Association Rule Extraction (knowledge discovery), Sequence Detection. Data Mining Methods may also be classified according to various criteria as: Decision Trees, Rule Induction, Neural Networks, Nearest Neighbor, Genetic Algorithms, Regression

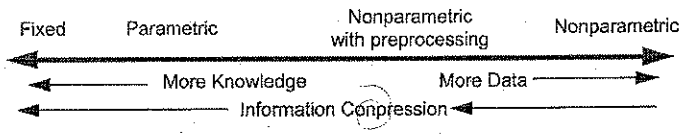

Fig. 1. The Modeling Method Continuum: Fixed models use existing knowledge on a problem(suc as in engineering). The nonparametric method relies on a large date set but does not use existing knowledge. The less-well-known aspect of a problem is captured by the nonparametric model.

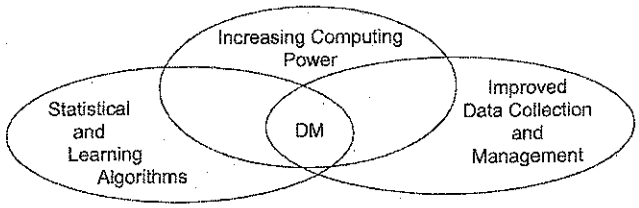

Fig. 2. Datamining Algorithms Development.

Models, Bayesianism, etc. At present datamining sits at the intersection of three broad and converging trends as shown in Fig. 2.

The Problem Space consists of: *Input dimensionality::* the number of components of the input vector; *Input space::* the set of allowed input vectors (typically infinite);*Mapping::* the model; the function that transforms/maps the inputs to the output

$$\vec{y} = f(\vec{x}) \tag{1}$$

where

$$\vec{x} = [x_1, x_2, \ldots, x_n]^T \tag{2}$$

and

$$\vec{y} = [y_1, y_2, \ldots, y_m]^T; \tag{3}$$

*Parameter vector::*a more accurate model is

$$\vec{y} = f(\vec{x}|\vec{\Theta}) \tag{4}$$

where $\vec{\Theta}$ is the parameter vector; and the *Learning algorithm::* generally supervised or nonsupervised learning, which fine-tune the parameters which are a part of the model.

Typically the basis of all datamining is some kind of a *clustering technique* which may serve as a preprocessing, and data reduction technique which may be followed by other algorithms for rule extraction, so that the data can be interpreted for and comprehended by humans. Prediction and classification may be a goal of the process also.

3 Clustering:: e.g. Patterns (Categorization)–Low Level Artificial Science

Clustering is the process of grouping data into classes or clusters so that objects within a cluster have high similarity in comparison with one another, but are very dissimilar to objects in other clusters. *Dissimilarities* are assessed based on the attribute values describing the objects. Often *distance measures* are used. Clustering is an unsupervised activity, or should be. Clustering can

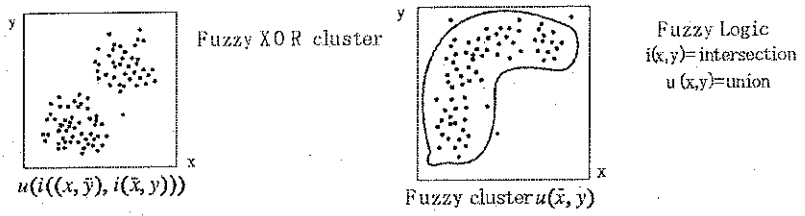

Fig. 3. 'Nonlinear' Cluster formation. The XOR kind of cluster is not linearly separable, and thus was not 'learnable' by the perceptron. This was pointed out by Minsky in the early days of neural networks and halted further research until Hopfield nets. The arbitrary shape clusters such as those above are still a problem for datamining algorithms. See for example, [7] for examples of nonlinear clustering.

be thought of as the preprocessing stage for much of datamining. An automated clustering algorithm may be said to be the goal of datamining since *classification* and *prediction* algorithms can work on the clusters. Similarly *association rules* may be derived from the clusters. Clustering can be used by marketers to discover distinct groups of buyers in their customer bases. It can be used to derive taxonomies in biology or linguistics. It can help categorize genes with similar functionality, and classify WWW documents for information discovery. In other words, it is a tool to gain insight into the distribution of data. It has been a branch of statistics for years. In machine learning it is an example of unsupervised learning. In datamining active themes for research focus on scalability of clustering methods, the effectiveness of methods for clustering of complex shapes, and types of data, high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases. Requirements for an ideal clustering procedure can be found in [6], and these capabilities can be inferred to exist in human beings, where they are performed by neural networks.

The memory-based clustering procedures typically operate on one of two data structures: data matrix or dissimilarity matrix. Every object is a vector of attributes, and the attributes may be on various scales such as nominal, ordinal, interval or ratio. The d(j,k) in the *dissimilarity matrix* is the difference (or perceptual distance) between objects j and k. Therefore d(j,k) is zero if the objects are identical and small if they are similar.

4 Mathematics of Belief (Cox-Jaynes Axioms)

We are at a point where we can begin to make sense of what science is, how it is to be done, and how reasoning takes place. Intelligence (of our kind, e.g. human kind) is embodied in the brain. The mind is what the brain does. Sensory impressions (data) are always coming in. The brain/ mind is always categorizing/recognizing things. It is when it starts to fail that we start to

Table 1.

Scalability	The procedure should be able to handle large number of objects, or should have a complexity of O(n), O(log(n)), O(nlog(n)). Human sensory systems seem to handle inputs logarithmically.
Ability to deal with different types of attributes	The method should be able to handle various levels of measurement such as nominal (binary, or categorical), ordinal, interval, and ratio. Infants start to learn to categorize objects (implicitly using measurement), then learn to name them, then we teach them about arithmetic, and formal reasoning (mathematics) later. Theoretical work in datamining is about the mathematics (formality) of how the human brain accomplishes these tasks.
Discovery of clusters with arbitrary shape	The procedure should be able to cluster shapes other than spheroidal which is what most distance metrics such as the Euclidean or Manhattan metrics produce. Humans can recognize all kinds of patterns. (See Fig. 3)
Minimal requirements for domain knowledge to determine input parameters	The method should not require the user to input various "magic parameters". Human infants, indeed all living things start learning from the environment immediately. Our (scientific,formal, mathematical) knowledge of different domains comes after many years of study.
Ability to deal with noisy data	The method should be able to deal with outliers, missing data, or erroneous data. Certain techniques such as ANNs seem better than others. Over hundreds of thousands of years we have been able to finally get a clear view of nature, via the scientific method, despite all the noise in the signal.
Insensitivity to the order of input records	The same set of data presented in different orderings should not produce different sets of clusters. Unfortunately, this does not seem to be true for humans; what we learn first affects our learning because of brain plasticity.
High dimensionality	Human eyes are good at clustering low-dimensional (2D or 3D) data but clustering procedures should work on very high dimensional data. The extension to higher dimensions cannot be done without formal mathematical models and training.
Constraint-based clustering	The procedure should be able to handle various constraints. Here, too, humans seem to excel; we have been able to find patterns in life and "use" them taking into account the constraints imposed by laws of nature.
Interpretability and usability	For practical purposes (bare minimum, for most people) this means that the results such as association rules should be given in terms of logic, Boolean algebra, probability theory or fuzzy logic. Any mathematics we can use is by definition 'comprehensible' in some sense.

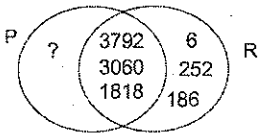

Fig. 4. Suppose we enter the integers 3792, 3060, and 1818. What do we want? What would you predict? How do we generalize? How do we learn (e.g.) via examples?

realize that what is apparently true (for the brain/mind) may not be true in reality. The same problem occurs in science; what folk physics says may not be what physics says. This interplay between appearances and reality have been the bane of philosophers since Plato's metaphor with the cave, and are the stuff of optical illusions psychologists study even today. Forced binary discrimination tests are conducted by psychologists to try to analyze perception at the detail level and similar ideas are used there. In this section we will look at the use of crisp logic, fuzzy logic and Bayesian reasoning.

At one time IBM was using a database system and method of searching called query- by-example (QBE), in an attempt to teach people natural ways of searching databases. QBE is both a query language and the name of a DB system including it. Something similar occurs when a child learns a language or when we search Google by entering some words or phrases. Google 'predicts' what is meant (e.g. appearence) and tries to make it match reality (what we really want). In real life we may never completely know reality. That is also the quandary that science finds itself in. In science, as in logic, truth is provisional; only falsehood is definite. Datamining is the quintessential kernel of science in that it tries to group data into clusters so that each member in the cluster resembles other members of the cluster more than it resembles other objects. That is what categorization is about, the lowest level of measurement. Even animals are capable of clustering and categorization. Much of this is done implicitly and informally in most living organisms. Such informality is unforgiveable in doing science even if not in writing for the masses.

Suppose, P is what Google predicts in response to a query by a user who really wants all the articles in the set R (reality). Then the cases (see Table 2) 00 and 11 are correct; and the cases 01 and 10 are errors. The case 00 corresponds to all articles the user did not want and did not get; 11 is the intersection of P and R, and consists of webpages the user wanted and got. Suppose we give an HIV test to millions of people and P is the set of people whom the test "predicted" have HIV; after all no test is perfect. The cases 00 and 11 are interpreted as above. The case 01 is called False Negative (FN) because the test predicts that the subject is negative but is not. Similarly 10 is False Positive (FP) because the test claims that the subject is infected but is not. Which error is more costly to society?

Similar conundrums bedevil society at all levels of decision-making.

Suppose these sets represent the people that the justice system claims are criminals, say murderers. Which error is worse? With all due respect to

Justice Holmes, a justice system that lets go 1000 murderers (FN) for every single FP is not a good system at all. Suppose the people being let go are terrorists bent on destruction. Is it a good idea to let 1000 of them escape? The same decision problem plagues police officers every day. The same decision problem plagues common people every day. People must make judgements based on appearances every day, and pay for the FPs and FNs every day. So it is especially with relationships, friendships, and beliefs about society. Shall we allow all kinds of harmful beliefs (say, Marxism of the worst kind) be propagated by the alleged elite at our universities subsidized by our tax dollars. Is this kind of decision-making not the essence of science? We have to make a mental shift from certain knowledge (things like $1 + 1 = 2$) to probabilistic truths. Many things which we may believe are true may not be true.

It is possible to treat belief systems with the rigor of mathematics. We can reason about reasoning. The result is the Cox-Jaynes axioms. The explanation below closely follows the exposition in Baldi & Brunak (1998). A hypothesis H about the world is a proposition, albeit a very complex one composed of many more elementary propositions. A model M can be viewed as a proposition. In the nonmathematical world the model M would be called a "theory". Since models have many parameters we may consider M=M(w) where w is a vector of parameters. We really wish to reason in the face of uncertainty. Thus we consider that given a certain amount of information I, we can associate with each M a degree of plausability or confidence, or degree of belief (DoB), B(X|I). Now, for any two propositions X, and Y either we believe X more than Y or vice versa or we believe both equally. Therefore we write

B(X|)>B(Y|I)

if we think that X is more plausable than Y. It is clear that the relationship (>) should be transitive. Formally, this is Cox's first axiom:

$$B(X|I) > B(Y|I) \wedge B(Y|I) > B(Z|I) \Rightarrow B(X|I) > B(Z|I) \qquad (5)$$

This means that DoBs can be expressed as real numbers and thus ">" is an ordering relationship. Therefore B(X|I) represents a real number, even if such a number is not easy calculate. For the next axiom consider the belief in the proposition B($\bar{X}$|I) Where $\bar{X}$ is the denial of the hypothesis. It is easy to see that the more confidence we have in B($\bar{X}$|I) the less confidence we should have that the denial B($\bar{X}$|I) is true. This belief of ours should be true for all such statements. Thus, in mathematical terms, we say that there exists a function so that

$$B(\bar{X}|I = f(B(X|I)) \qquad (6)$$

Notice that we have not given the form of the function F that specifies how the two propositions should be decreasing functions of each other, merely that they should be related to one another somehow. Certainly, it makes sense that as we believe X more and more, we should believe NOT(X) less and less. Later, how this decreasing function should be incorporated into the

Cox- Jaynes axioms will be made explicit, The third axiom considers pairs of hypotheses;

$$B(X \bullet Y|I) = g(B(X|I), B(Y|X, I)) \tag{7}$$

Where the • sign indicates a logical-AND. In other words, our degree of belief that X is true, and our belief that Y is true, for example, depends on our belief that X is true, and that Y is true knowing that X is true. Now, I is the conjunction of all the available pieces of information. It can represent background knowledge; it can include specific experimental data or other data. When it is necessary to be specific we can write I=I(D) to denote dependence on a corpus of data D. It is not really fixed and can be augmented by other symbols or even dropped when it is well-defined. These three axioms determine, up to a scaling, how to calculate DoBs. It can be proven that there is always a scaling of degrees of belief such that DoBs can be constrained to [0,1] e.g. a function P(X|I)=k(B(X|I))

Furthermore it can be shown that P(X|I) is unique and that is satisfies all the rules of probability. Specifically, if *DoB* is restricted to the interval [0,1] then the functions F and G must be given by f(x) = 1-x and g(x,y) = xy.

In this case, the DoBs can be replaced by probabilities. It should be noted that f(x) is really like the fuzzy-negation, and g(x,y)=xy corresponds (in some fuzzy logics) to a logical-AND or intersection. As a result the second axiom can be written as the sum rule of probability theory e.g.

$$P(X|I) + P(\bar{X}|I) = 1 \tag{8}$$

This is simply the probabilistic equivalent of the Law of the Excluded Middle in logic that goes back to Aristotle. The third axiom is the product rule, e.g.

$$P(X \bullet Y|I) = P(X|I)P(Y|X \bullet I) \tag{9}$$

These DoBs can simply be replaced by probabilities. By using the symmetry one can then obtain the important Bayes theorem

$$P(X|Y \bullet I) = \frac{P(Y|X \bullet I)P(X|I)}{P(X|I)} = P(X|I) \cdot \frac{P(Y|X \bullet I)}{P(Y|I)} \tag{10}$$

This rule implies inference-learning since it describes how to update our degree of belief P(X|I) in X, in light of the new pieces of information provided by Y, to obtain the new P(X|Y • I) Because of this P(X|I) is called the prior probability and P(X|Y • I) the posterior probability. Obviously the rule can be iterated as many times as needed as new information becomes available. It should be noted that there is a more general set of axioms for a more complete theory that encompasses Bayesian theory. These are the axioms of decision theory or utility theory which focuses on how to make optimal decisions in the face of uncertainty. The simple axioms of decision theory allow the construction and estimation of Bayesian probabilities. An even more general theory, game theory, considers the case where the uncertain environment includes

other intelligent agents or players. In dealing with nature, we assume that nature does not change its tactics to oppose us as we discover its laws hence Bayesianism is more or less sufficient. In a more specific setting, we are most interested in deriving a parameterized model M=M(w) from a corpus of data D. Dropping I for simplicity, the Bayes theorem gives.

$$P(M|D) = P(M) \cdot \frac{P(D|M)}{P(D)} \tag{11}$$

The prior P(M) is our estimate of the probability that model M is correct before we have obtained any data. The posterior P(M|D) represents our updated belief in the probability that model M is correct given that we have observed the data D. Indeed, this is more or less what doing science is about. There is more on this topic, namely being able to compare models to one another such as using the Akaike information criteria (AIC), or MDL (minimum description length) or EM (expectation maximization); these can be found in books on datamining, and artificial intelligence. It is easy to see that M is appearance/prediction (e.g. our model/prediction of reality) and that D represents (information/data or reality). We now have a way of reasoning precisely using beliefs, probabilities, logic and fuzzy logic, and creating simple (fuzzy-logical) models of various aspects of reality.

5 Hempel's Problem Revisited with Measurement

Hempel's Raven paradox involves the 'confirmation value' of empirical evidence in the logic of epistemology. The statement "Ravens are black" (e.g. R $\Rightarrow$ B) via the contrapositive is "equivalent" in logic to "If x is not a raven, it is not black" (e.g. $\bar{\mathrm{B}} \Rightarrow \bar{\mathrm{R}}$). Therefore, if a raven that is black "confirms" the statement R $\Rightarrow$ B, then a red banana confirms the same statement via the contrapositive. These are the problems one runs into when attempting to do probability theory via binary logic. There are four possibilities when faced with a perception/cognition problem or a generalization problem (see Fig. 5 or Fig. 6 below). Perception/cognition, typically as discussed in philosophy and psychology consists of 'naming' objects e.g. categorization. The simplest form of it is the "Forced Binary Discrimination Test" often used in psychology to analyze cognition. This process itself can be explained in terms of a nonlinear differential equation [8]. "Truth" (degree of truth?) depends on the relative sizes of the various sets shown in Fig. 6. We are back to Plato's Problem (e.g. Fig. 5).

Clearly, the regions marked 01 and 10 are errors where the appearance does not match reality. The case 01 corresponds to black nonraven things which are weak confirmation. The case 10 corresponds to ravens that are not black and is the counterexample to the generalization. Certainly this is definite nonconfirmation. We want region 11 to be high because that confirms

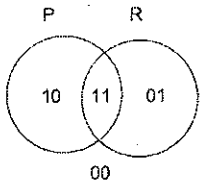

Table 1:

Appearance	Reality	Judgement	
0	0	Good	
0	1	Type I Error	False Negative (FN)
1	0	Type II Error	False Positive (FP)
1	1	Good	

Fig. 5. Plato's Problem: Appearance vs Reality. How science is done using simple crisp logic and sets. The terms False Negative and False Positive are especially useful in medical testing.

Fig. 6. Here α is used for "apparent" and ρ for "real" in order to prepare for developments below.

REALITY / APPEARANCE	0 negative	1 Positive
Negative 0	B(*Apparently False* | *Really False*)	B(*Apparently False* | *Really True*)
Positive 1	B(*Apparently True* | *Really False*)	B(*Apparently True* | *Really True*)

Fig. 7. Belief (probability) of Appearance and Reality.

the generalization. In Fig. 6 we see the general sizes of the various categories that we expect in the real world. The type 00 is the largest category and its confirmatory value is almost zero. Then as an initial starting point we should want 11/01 to be high and 10/00 to be high. Therefore we should try to maximize something like (11/01)(10/00). Or we might want to deal with the fractions (11/ [11+01]) and (10/[10+00]). The former is the fraction that is correct[true] of the total correct[true]. The second is the fraction that is incorrect[false] out of the total that is incorrect[false]. Can we change our reasoning so that we can make better decisions?

For normal reasoning these do not obey things like the laws of probability theory because they are beliefs and we cannot make these computations on the fly. So we can define

$$\tau = \frac{B(\alpha|\rho)}{B(\alpha|\rho) + B(\bar{\alpha}|\rho)} = \frac{1}{1 + \frac{B(\bar{\alpha}|\rho)}{B(\alpha|\rho)}} = True\ BeliefFraction \tag{12}$$

$$\Phi = \frac{B(\alpha|\bar{\rho})}{B(\bar{\alpha}|\bar{\rho}) + B(\alpha|\bar{\rho})} = \frac{1}{1 + \frac{B(\bar{\alpha}|\bar{\rho})}{B(\alpha|\bar{\rho})}} = False\ BeliefFraction \tag{13}$$

But the fraction τ is $\frac{11}{11+01}$ or $\frac{1}{1+01/11}$ and the fraction Φ is $\frac{10}{00+10}$ or $\frac{1}{1+00/10}$. Minimizing (01/11) will maximize τ and maximizing (00/10) will minimize Φ. Therefore, minimizing (01/11)*(10/00) will optimize τ/Φ which is what we want, if the costs are equal. If not we should take into account costs. But there is already a developed theory based on conditional probabilities using the Bayes Theorem. The True Belief fraction is called sensitivity, and the False Belief Fraction is 1-specificity. These terms are from medical terminology of FN and FP.

6 Hypothesis Testing and Receiver Operating Characteristics (ROC)

Suppose some event has occurrred. We have to decide whether it is important (valuable opportunity e.g. "chance" for great gain of some sorts) or that it is worthless. It corresponds to making a decision as to whether something is real or fake (an impostor opportunity). Given some kind of a score, or measurement, classification (into real truth vs apparent truth, etc) involves choosing between two hypotheses: that the apparent oppportunity is real, or fake. Let H0 be the hypothesis that the chance is fake, and H1 be the hypothesis that the chance is real. Suppose we are given the measurements from two different pdfs according to whether the chance is real or fake as shown in Fig. 8.

In Table 2 we see the various possibilities. It is clear that the True Belief Fraction is really the sensitivity of the test, and the False Belief Fraction is the FPR (false positive rate). Table 2 summarizes the possibilities.

The hypothesis determines over which pdf to integrate, and the threshhold, T, determines which decision region forms the limits of integration. Let $p(z|H_0)$ be the conditional density function of the score z generated by false-chance events, and let $p(z|H_1)$ be the conditional pdf for true/ real chance events. If the true conditional pdfs of fake and real chances are known, then

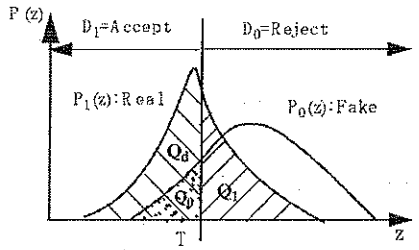

Fig. 8. Statistical Inferencing Scheme for ROC The probability density functions of real and fake (false) chances or opportunities for gain of some kind. (see Campbell or Metz).

Table 2.

Performance Probabilities	Decision D	Hypothesis H	Name of Probability	Result	Explanation
Q_0	1	0	Size of Test "Significance"	Type I Error	FP False Acceptance
Q_1	0	1		Type II Error	FN False Rejection
Q_d=1-Q_1	1	1	Power of Test		True Acceptance
1-Q_0	0	0			True Rejection

the Bayes test with equal misclassification costs is based upon the likelihood ratio for the decision- maker A, $\lambda_A(z)$

$$\lambda_A(z) = \frac{p_A(z|H_0)}{p_A(z|H_1)} \tag{14}$$

The probability of error, which is minimized by Bayes' decision rule is determined by the amount of overlap in the two pdfs. The smaller the overlap between the two pdfs, the smaller the probability of error. If the true conditional pdfs score densities for the good-decision-maker vs the bad- decision-maker are unknown, the two pdfs can be estimated from sample experimental outcomes. The conditional pdf given a good decision-maker A, $p_A(z|H_1)$ can be estimated from the decision-maker's own scores using his model. The conditional pdf for bad-decision-makers, $p_A(z|H_0)$ is estimated from other decision-makers' scores using the A's model. Once the likelihood ratio for A, $\lambda_A(z)$ can be determined, the classification problem can be stated as choosing a threshold T so that the decision rule is

$$\lambda_A(z) \begin{cases} \geq T & \text{choose } H_0 \\ < T & \text{choose } H_1 \end{cases} \tag{15}$$

One of the ways in which the threshold T can be determined is by varying T to find different FA/ FR (FP/FN) ratios and choosing T to give the desired FP/FN ratio [2]. Since the either of the two types of errors can be reduced at the expense of an increase in the other, a measure of overall system performance must specify the levels of both types of errors. The tradeoff between FP and FN is a function of the threshold. This is depicted in the ROC (receiver operating characteristic) curve. In the example by Campbell, it is assumed that the product of the probability of FP and the probability of FN is a constant for this system (which is not true in general) and is equal to the square of the equal error rate (EER). In the example by Metz, the ROC curve is the one in which the sensitivity is plotted against the false positive fraction which is 1-specificity. It comes from the area of statistical analysis, specifically, medical testing. A test is given to determine if a person is sick with a given disease; such a person is said to be positive. These are clearly related to the False-Positive (FP) and False-Negative (FN) results of statistical testing. Now, it is not clear at all that tests give constant results, or that FP is linearly related to the FN. Specifically, in medicine, Sensitivity and Specificity of each test depend on the particular "threshhold of abnormality" adopted for that test. Some tests have higher Sensitivity but lower Specificity than the other. Perhaps we should maximize specificitysensitivity or something like it.

With tests we are stuck with what tests can do. Similar comments can be made about psychological tests, and informal judgements people make which are based on heuristics. The reason the above analysis is not sufficient is that there will be built in biases against and for truth vs falseness. For example, there are loss aversion bias, status quo bias, hindsight bias, bias towards positive and confirming evidence, and others which have been studied [see for example, [10]]. These biases will be a function of socialization and knowledge capability. In the same paper a simulated example is given of the computations for various thresholds as shown in Fig. 9.

We can see that... the cost itself can be treated as a fuzzy variable so that the final result resembles something like a an fuzzy XOR. There are many

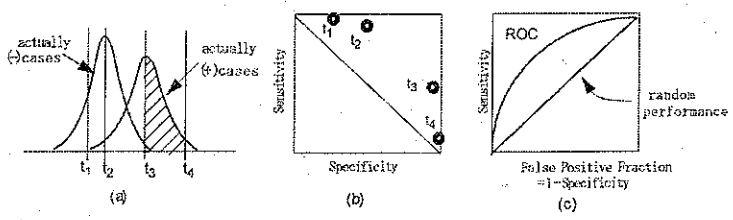

Fig. 9. Simulated Results from Metz. In (a) we see four simulated positions for the 'anomaly'. The resulting points on the plot of sensitivity vs specificity is shown in (b). A curve is swept out as the threshold of abnormality (T) is varied continuously (c) and compared against "random" guessing (straight line).

other kinds of problems that resemble this. Consider now the interpretation of Eq. (12a) which describes, say, results of a forced binary test

$$\frac{P}{1-P} = \tau \tag{16}$$

where $\tau = K/D$, P=Pr(correct answer), K=knowledge, and D=problem difficulty. This result can be considered to be a special form of Rasch modeling, which is given by

$$\pi = \frac{e^{\lambda}}{1+e^{\lambda}}$$

or

$$\frac{\pi}{1-\pi} = e^{\lambda}$$

where

$$\lambda = K - D$$

and

$$\pi = \Pr(\text{correct answer}) \tag{17}$$

For a simple case such as a binary test, P can easily be interpreted as the TPR, and $1-P$ is the FPR. Consider the case $\tau = 0$ (when the knowledge of the individual is 0 or if the problem is of infinite difficulty), then clearly $P = 0$. Now, if the difficulty is 0, then we consider the equation $\{1-P\}/P = 1/\tau$. Clearly, then $1 - P = 0$ or $P = 1$. For an interpretation of $K = D$ consider the case of a match between two boxers of exactly equal capability. The probability of either of them winning should be 0.5, and this is the case with this equation. Therefore it is clear that, for someone randomly guessing at the solution (binary choice) we'd expect a linear relationship between τ and Φ,

$$\tau = f(K/D)\Phi = f(\kappa)\Phi = \kappa\Phi \tag{18}$$

Anything better than random guessing would be a curve above the straight line at 45 degrees and anything below would be worse than random.We can derive the typical ROC curve (which is derived via some assumptions) by assuming different values of κ and therefore something like a piecewise linear relationship between τ and Φ. Specifically, suppose sets of questions at different difficulty levels were given to students. For each set of questions we'd expect different values of τ and Φ. Since we know the difficulty levels, we can compute the corresponding K values (see Fig. 10). Now for a multiple choice test, we can apply the same interpretation. Indeed, this would be the ideal way to grade tests. Someone operating in, say, a business environment making decisions at a technical level as to whether a given development in the field is a great "chance" with a potential bonanza for the company would be faced

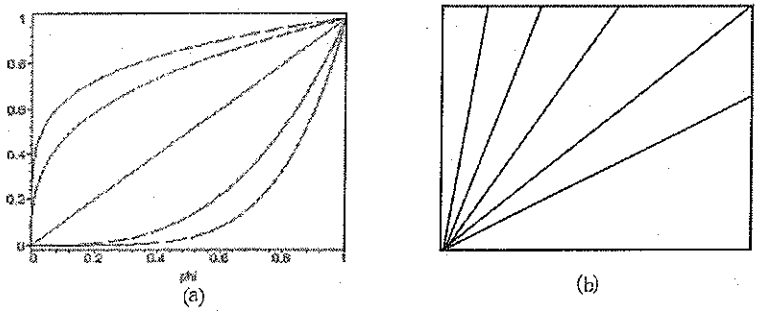

Fig. 10. The curve given by for various values in $\tau = f(k)\Phi^{1/f(k)}$. For values of $1/f(k) > 1$ we obtain the curves below the random guessing line. For a linear relationship e.g. $\tau = k\Phi$, obviously we get the straight lines in (b). These must be considered to be linear approximations of a more complex nonlinear reality.These curves can be related to learning theory (see below).

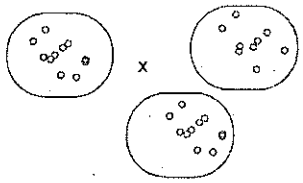

Fig. 11. Chance Discovery Problem and [In]formal Datamining.

with a similar problem. We can designate the function $f(\kappa)$ as the knowledge function, or the competence function. For example, via theoretical reasons we may assume other functional forms such as

$$\tau = f(\kappa)\Phi^{1/f(\kappa)} \tag{19}$$

These latter ones will result in curves as shown in Fig. 10.

7 Categorical Measurement

There are errors in measurement and judgement. There could be confusion in perception and categorization so that only nominal (categorical level) of measurement may be possible and the person making that decision may make errors in judgement. The person might hold inconsistent beliefs e.g. X +Not(X) is not equal to 1. There may be biases in belief e.g. folk-belief may interfere totally with Cox-Jaynes axioms. Here we cannot find a measurment along a single dimension so that the ROC curves explained on the bass of the threshold T, may not be possible. A picture (Euler diagram, as in Fig. 11) can be used to explain what we might be able to do in such a case. A

discovery, (or fact/information) X, could possibly have big consequences (e.g. a chance). Its evaluation (apparent value) depends on the knowledge of the discoverer, but its real value is not known. It can be miscategorized. It might be fit into one of the clusters (say, a branch of mathematics), or it may be used to connect the two branches or even all three areas into a complete more comprehensive theory. This process is a part of the continuous generation and production of [scientific] knowledge.

Suppose the values possible are: worthless (W), valuable (V), and priceless(P). These would correspond in Fig. 11 to (i) not connecting X to anything, (ii) connecting X to one of the clusters, and (iii) connecting X to all three clusters and thus joining all three fields together in a new theoretical finding. These 'fields' may be cross-disciplinary or in a business setting they may cross the boundaries of conventionally different branches of business.It is possible to convert this into something like the Likert Scale (and use ordinality) and therefore relate this to the concept of a threshold as above, but here we leave the categories as they are and make assumptions about the capabilities of the reasoner.

The analysis below closely follows that of [14]. The overall probabilities with which the various possibilities (e.g. chances) are judged are π_W, π_V, and π_P. The probability that a true possibility of type T is believed as a possibility of type t is denoted as R(T|t) (e.g. R(real|apparent). We set this equal to $1-\gamma$ for a correct observation, and $\gamma/2$ for either of the possible wrong judgments. The idea can be used for multiple-choice tests. For example, the probability that a valuable possibility will appear to be valuable is R(V, V) = 1 - γ, and the probability that a valuable possibility will appear to be worthless is $R(W, V) = 1 - \gamma/2$. The probability P(T|t) that the true type is T when the observed (apparent) type is t is P(T|t)=P(T_t)/P(t). For example, the probability that a type which appears to be worthless is in reality worthless is

$$P(W|W) = \frac{R(W|W)\pi_W}{R(W|W)\pi_W + R(W|V)\pi_V + R(W|P)\pi_P}$$

$$= \frac{(1-\gamma)\pi_W}{(1-\gamma)\pi_W + (\gamma/2)\pi_V + (\gamma/2)\pi_P} \tag{20}$$

The probability that the chance which appears to be priceless (extremely valuable) opportunity is in reality valuable is given by

$$P(P|W) = \frac{R(P|W)\pi_W}{R(P|W)\pi_W + R(P|V)\pi_V + R(P|P)\pi_P}$$

$$= \frac{(\gamma/2)\pi_W}{(1-\gamma)\pi_W + (\gamma/2)\pi_V + (\gamma/2)\pi_P} \tag{21}$$

The case for N categories is simply a generalization of these equations

$$P(C_n|C_m) = \frac{R(C_n|C_m)\pi_{C_m}}{\sum_m R(C_n|C_m)\pi_{C_m}} \tag{22}$$

It should be recalled that for each decision maker the probability of making the correct identification is fixed at $R(C_k, C_k) = 1 - \gamma$, and the probability of making the wrong identification is distributed equally so as to maximize the entropy. In cases in which there is some dimension along which some measurement(s) can be made these probabilities could be distributed according to a pdf along this dimension, which would then make this more like the ROC; for this we would need a concept of distance (see below).

8 Summary

1) The classical ROC (the one shown above) is based on signal detection, and thus the Gaussian assumption. The integral leads to solutions in terms of the error function, erf(T), or erfc(T) so that explicit (and simple) algebraic equations are not possible. It would be better in many ways if simpler equations were used, if for no other reason than pedagogical purposes. For example, a uniform density (which would be better in some cases) would produce easily comprehensible algebraic equations. A triangular density (an approximation to Gaussian) would also produce easily manipulatable algebraic quantities. This is shown in Fig. 13.
2) The ROC method as shown actually hides the fact that the relative positions of the error pdf and the true signal pdf are fixed (e.g. to the right or left). In reality, in complex situations, even if presented within the ROC framework, the decision-maker has to make a decision as to the relative positions of these pdfs and that itself is not a part of the ROC method except implicitly.
3) With some simplifying assumptions the categoricity of the last section can be turned into the threshold framework (with ordinal, difference or ratio levels of measurement) of the 'classical' ROC shown earlier. The main problem is that the "distance" measure cannot be represented along a single dimension as in the graphs shown. Instead, we might have to use something like a graph-theoretic measure on complete graphs. For example, the Likert scale would require a K5 (Kuratowski graph) complete graph of 5 nodes. Each incorrect edge (4 of them) would/could be given a weight (distance) of $1/\gamma$. If we tried to use the ROC analysis as above, we'd have to (i) index the categories and create

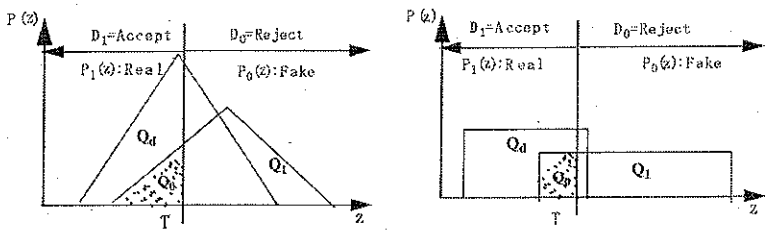

Fig. 12. Pedagogical Simplification of the ROC Analysis.

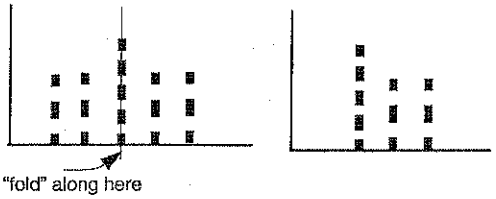

Fig. 13. Categorical Bayes and Distance.

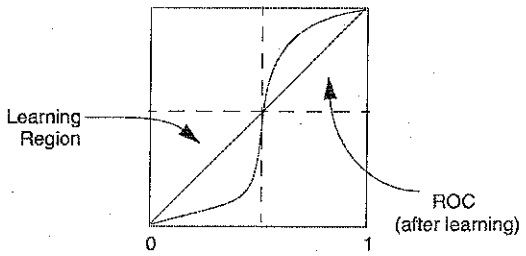

Fig. 14. ROC and Learning Theory. It should be noticed that the abcissa is a 'normalized T' and thus the classical Verhulst curve will not work.

some kind of a distance/similarity measure, or (ii) use a "folded" distance e.g. Fig. 13.

4) The ROC method itself should be a part of learning theory. See Fig. 14.

During learning where the learner has many choices, there's a high probability of false positives since we learn from errors (negative feedback). This is the part of the curve in the first quadrant. Obviously the axes would have to be normalized so that the Verhulst curve would have to be fixed up (e.g. normalized). Ditto for the related Rasch curve which is specifically related to measurement of competence. And this would also be used in testing in both a school setting and in informal testing of competence in various fields especially of artificial agents since they do not suffer from psychological needs.

5) A more sophisticated (future) approach would extend this so that 'knowledge' can be defined more clearly, especially domain knowledge that is required to make good decisions. Only hints have been given here. Obviously, it would be most easily done in the case of scientific (mathematical) knowledge. In cases where information is significant, multiple choice questions would be used.

6) The ROC results can easily be explained in terms of fuzzy-logic. Even if for no other purposes than pedagogical, it would be extremely useful to so do. Indeed some of the equations above are easily explicable and comprehensible in terms of fuzzy logic. Sensitivity and specificity are basically complements.

Plotting both against normalized-T would lead to two curves resembling supply and demand curves of economics (e.g. tradeoff). The normalized product basically the Beta density. If the costs of false positives and false negatives are equal the function is symmetric and can be considered the analog of the Gaussian in normalized [0,1] space. If the costs are not equal, the result is skewed in one of the directions. In either case, the function can be treated as an *fdf* (fuzzy density function). In other words, the costs of correct and incorrect decisions have not been incorporated. They can also be incorporated so as to depict the problem as an optimization problem which would be more appropriate for a business or economic setting, and since ultimately analysis has to be justified in terms of economic costs and benefits, this line of research should be pursued.

References

1. Baldi, Pierre and Soren Brunak (1998), Bioinformatics:The Machine Learning Approach, MIT Press, Cambridge.
2. Campbell, J. (1997) Speaker Recognition: A Tutorial, Proceedings of the IEEE, Vol 85, No. 9, September 1997, pp. 1437-1462
3. Fawcett, T. (2004) ROC Graphs: Notes and Practical Considerations for Researchers, http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf
4. Flach, P., The many faces of ROC analysis in machine learning, www.cs.bris.ac.uk/ flach
5. Godfrey-Smith, P (2003) Theory and Reality: an introduction to the philosophy of science, University of Chicago Press, Chicago.
6. Han, Jiawei and Micheline Kanber, (2000) "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers.
7. Hubey, H.M.(2005) Detection of Chance Events, Real-time Processing of High-Dimensional Streaming Data, and Datamining, in Readings in Chance Discovery, eds A. Abe and Y. Ohsawa, Advanced Knowledge International, Adelaide, Australia, 2005.
8. Hubey, H.M., (1999) Mathematical and Computational Linguistics, Lincom Europa, Munchen, Germany.
9. Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian netoworks, Volume 19, Number 17, November 22, 2003, pp.2271-2282, Electronic Edition http://bioinformatics.oupjournals.org
10. Kahnemann, D., P. Slovic, and A. Tversky (Eds.) (1982). Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press.
11. Metz CE. Semin Nuclear Med 1978 VIII(4) 283-298. Basic principles of ROC analysis.
12. Ohsawa, Yukio, http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa/ChanceDiscovery.html
13. Park, Seong Ho, Jin Mo Goo, and Chan-Hee Jo,(2004) Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists, Korean J. Radiology 5(1), March 2004, pp.11-18.

14. Sozou, P. and R. Seymour (2005), Costly but worthless gifts facilitate courtship, Proceedings of the Royal Society, B (2005) 272, 1877-1884. Published online 26 July 2005.
15. Steingold, S. Datamining: Techniques and Challenges: http://www.podval.org/sds/data/dm/ps
16. Binary Choice ROC applet:: http://wise.cgu.edu/sdt/sdt.html