Knowledge Discovery in Large Image Databases: Dealing with Uncertainties in Ground Truth

Padhraic Smyth[†], Michael C. Burl[†‡], Usama M. Fayyad[†], and Pietro Perona[‡]

[†] Jet Propulsion Laboratory
California Institute of Technology
MS 525-3660 — Pasadena, CA 91109
{pjs,fayyad}@aig.jpl.nasa.gov

[‡] Electrical Engineering
California Institute of Technology
MS 116-81 — Pasadena, CA 91125
{burl,perona}@systems.caltech.edu

Abstract

This paper discusses the problem of knowledge discovery in image databases with particular focus on the issues which arise when absolute ground truth is not available. The problem of searching the Magellan image data set in order to automatically locate and catalog small volcanoes on the planet Venus is used as a case study. In the absence of calibrated ground truth, planetary scientists provide subjective estimates of ground truth based on visual inspection of Magellan images. The paper discusses issues which arise in terms of elicitation of subjective probabilistic opinion, learning from probabilistic labels, and effective evaluation of both scientist and algorithm performance in the absence of ground truth. Data from the Magellan volcano detection project is used to illustrate the various techniques which we have developed to handle these issues. The primary conclusion of the paper is that knowledge discovery methodologies can be modified to handle lack of absolute ground truth provided the sources of uncertainty in the data are carefully handled.

1 Introduction

At last year's KDD workshop we presented initial results on building an automated system for locating and cataloging small volcanoes on the surface of Venus, using radar images returned by the Magellan spacecraft [Fayy93]. In particular we discussed some of the issues which are unique to the problem of knowledge discovery in *image* databases. This paper tackles the problem of developing and evaluating knowledge discovery systems for image databases when ground truth is not available. Remote-sensing image analysis problems where there is a complete absence of ground truth are increasingly common: the Magellan volcano detection project is used as an illustrative example. Indeed, the problem is not unique to image analysis applications, it also arises in certain medical diagnosis applications where absolute diagnoses are never known with certainty [Henk90]. In general, one cannot be sure that class labels (or values of the "dependent variable") are noise-free. The core issue addressed in this paper is that, in some cases, this uncertainty in class labels, or "lack of ground truth", must be dealt with explicitly: it needs to be estimated and then it needs to be accounted for by the learning algorithms employed.

The phrase "lack of ground truth" requires some comment: what is typically available is not a complete lack of ground truth, but rather, subjective estimates of ground truth. In other words, a domain expert (or group of same) examines the available data (an image) and provides a subjective estimate of the class labels for particular locations within the image. Hence, there is an additional source of noise in the data, namely *the noisy estimates from the expert.* It is critical that this noise source be modeled and calibrated as far as possible. The alternative is to ignore the noisy nature of the labelling process, assume that the labels are correct, and condition all algorithm design, parameter estimation, and performance evaluation on this premise. If the labelling process is not very noisy this is often the practical approach.

In this paper we focus on the case where there is considerable visual ambiguity in the images, such that there will be significant differences on the same data between the labels of the same expert at different times and between different experts. Ignoring this source of noise is likely to lead to a significantly miscalibrated

system. For example, in the volcano detection problem, the local density of volcanoes in a given planetary region is a parameter of significant geological relevance. Ignoring the subjective uncertainty in the labelling would lead to a systematic bias in terms of over-estimating local volcano densities.

The paper is structured in the following manner: first the general background to the problem is described, namely the Magellan mission and the scientific importance and relevance of investigating volcanism on Venus. We then review our overall philosophy behind developing "user-trainable" tools for knowledge discovery in databases, focusing in particular on the development of machine learning and pattern recognition tools which allow a scientist to train a search algorithm based on sample objects of interest. This sets the stage for the main discussion of the paper: the modeling and treatment of subjective label information. We outline the experimental methodology and basic principles of subjective elicitation, using data obtained from the participating scientists as examples. The following issues are then discussed in some detail: noise models to relate probabilistic labels to ground truth, performance evaluation metrics which incorporate probabilistic labels, and learning algorithm modifications. We note that previous work in the pattern recognition literature has dealt with some of the general theoretical aspects of this problem [Lug92, Silver80]; the originality of the work described here lies in the handling of the ground truth ambiguity problem in the context of a large-scale, real-world, image analysis problem.

2 Background: Automated Detection of Volcanoes in Radar Images of Venus

Both in planetary science and astronomy, image analysis is often a strictly manual process and much investigative work is carried out using hardcopy photographs. However, due to the sheer enormity of the image databases currently being acquired, simple manual cataloging is no longer a practical consideration *if all of the available data is to be utilized.* The Magellan Venus data set is a typical instance of the now familiar data glut situation in scientific, medical, industrial and defense contexts.

The background to this work is the notion of a trainable image analysis system; a scientist trains the system to find certain geological features by giving it examples of features to be located. The scientist can thus customize the tool to search for one type of feature versus another simply by providing positive and negative examples. In addition to automating laborious and visually-intensive tasks, the system provides an objective, examinable, and repeatable process for detecting and classifying objects in images. This allows scientists to base their analysis results on uniformly consistent data, free from subjective variations that invariably creep in when a visually exhausting task requiring many months or years is undertaken.

The Magellan spacecraft transmitted back to earth a data set consisting of over 30,000 high resolution synthetic aperture radar (SAR) images of the Venusian surface. This data set is greater than that gathered by all previous planetary missions combined — planetary scientists are literally swamped by data [Fayy94]. The study of volcanic processes is essential to an understanding of the geological evolution of the planet [Guest92]. Central to volcanic studies is the cataloging of each volcano location and its size and characteristics. There are estimated to be on the order of 10^6 visible volcanoes scattered throughout the 30,000 images [Aubele90]. Furthermore, it has been estimated that manually locating all of these volcanoes would require on the order of 10 man-years of a planetary geologist's time to carry out.

Empirical results using spatial eigenrepresentations (combined with supervised classification algorithms) have demonstrated that a trainable image analysis system can be roughly competitive with humans in terms of classification accuracy [Burl94, Fayy94]. The system uses a matched filter (for example, the mean of locally windowed training examples of volcanoes) to initially focus attention on local regions of interest. The detected local regions are projected into a subspace consisting of significant principal components of the training data. Although the full covariance matrix of the data (whose largest eigenvectors correspond to principal components) cannot be computed given the available sample sizes, the approximate eigenvectors can be computed using the singular value decomposition (SVD) technique [Burl94]. Supervised learning is used to produce a model which can discriminate between volcano and non-volcano local regions in the projected subspace. A simple maximum-likelihood multi-dimensional Gaussian classifier with full covariance matrices has been found to perform as well as alternative non-parametric methods such as neural networks and decision trees for the problem of discriminative learning in the projected eigenspace [Burl94].

Knowledge Discovery in Large Image Databases: Dealing with Uncertainties in Ground Truth

Padhraic Smyth[†], Michael C. Burl[†‡], Usama M. Fayyad[†], and Pietro Perona[‡]

[†] Jet Propulsion Laboratory
California Institute of Technology
MS 525-3660 — Pasadena, CA 91109
{pjs,fayyad}@aig.jpl.nasa.gov

[‡] Electrical Engineering
California Institute of Technology
MS 116-81 — Pasadena, CA 91125
{burl,perona}@systems.caltech.edu

Abstract

This paper discusses the problem of knowledge discovery in image databases with particular focus on the issues which arise when absolute ground truth is not available. The problem of searching the Magellan image data set in order to automatically locate and catalog small volcanoes on the planet Venus is used as a case study. In the absence of calibrated ground truth, planetary scientists provide subjective estimates of ground truth based on visual inspection of Magellan images. The paper discusses issues which arise in terms of elicitation of subjective probabilistic opinion, learning from probabilistic labels, and effective evaluation of both scientist and algorithm performance in the absence of ground truth. Data from the Magellan volcano detection project is used to illustrate the various techniques which we have developed to handle these issues. The primary conclusion of the paper is that knowledge discovery methodologies can be modified to handle lack of absolute ground truth provided the sources of uncertainty in the data are carefully handled.

1 Introduction

At last year's KDD workshop we presented initial results on building an automated system for locating and cataloging small volcanoes on the surface of Venus, using radar images returned by the Magellan spacecraft [Fayy93]. In particular we discussed some of the issues which are unique to the problem of knowledge discovery in *image* databases. This paper tackles the problem of developing and evaluating knowledge discovery systems for image databases when ground truth is not available. Remote-sensing image analysis problems where there is a complete absence of ground truth are increasingly common: the Magellan volcano detection project is used as an illustrative example. Indeed, the problem is not unique to image analysis applications, it also arises in certain medical diagnosis applications where absolute diagnoses are never known with certainty [Henk90]. In general, one cannot be sure that class labels (or values of the "dependent variable") are noise-free. The core issue addressed in this paper is that, in some cases, this uncertainty in class labels, or "lack of ground truth", must be dealt with explicitly: it needs to be estimated and then it needs to be accounted for by the learning algorithms employed.

The phrase "lack of ground truth" requires some comment: what is typically available is not a complete lack of ground truth, but rather, subjective estimates of ground truth. In other words, a domain expert (or group of same) examines the available data (an image) and provides a subjective estimate of the class labels for particular locations within the image. Hence, there is an additional source of noise in the data, namely *the noisy estimates from the expert.* It is critical that this noise source be modeled and calibrated as far as possible. The alternative is to ignore the noisy nature of the labelling process, assume that the labels are correct, and condition all algorithm design, parameter estimation, and performance evaluation on this premise. If the labelling process is not very noisy this is often the practical approach.

In this paper we focus on the case where there is considerable visual ambiguity in the images, such that there will be significant differences on the same data between the labels of the same expert at different times and between different experts. Ignoring this source of noise is likely to lead to a significantly miscalibrated

system. For example, in the volcano detection problem, the local density of volcanoes in a given planetary region is a parameter of significant geological relevance. Ignoring the subjective uncertainty in the labelling would lead to a systematic bias in terms of over-estimating local volcano densities.

The paper is structured in the following manner: first the general background to the problem is described, namely the Magellan mission and the scientific importance and relevance of investigating volcanism on Venus. We then review our overall philosophy behind developing "user-trainable" tools for knowledge discovery in databases, focusing in particular on the development of machine learning and pattern recognition tools which allow a scientist to train a search algorithm based on sample objects of interest. This sets the stage for the main discussion of the paper: the modeling and treatment of subjective label information. We outline the experimental methodology and basic principles of subjective elicitation, using data obtained from the participating scientists as examples. The following issues are then discussed in some detail: noise models to relate probabilistic labels to ground truth, performance evaluation metrics which incorporate probabilistic labels, and learning algorithm modifications. We note that previous work in the pattern recognition literature has dealt with some of the general theoretical aspects of this problem [Lug92, Silver80]; the originality of the work described here lies in the handling of the ground truth ambiguity problem in the context of a large-scale, real-world, image analysis problem.

2 Background: Automated Detection of Volcanoes in Radar Images of Venus

Both in planetary science and astronomy, image analysis is often a strictly manual process and much investigative work is carried out using hardcopy photographs. However, due to the sheer enormity of the image databases currently being acquired, simple manual cataloging is no longer a practical consideration *if all of the available data is to be utilized.* The Magellan Venus data set is a typical instance of the now familiar data glut situation in scientific, medical, industrial and defense contexts.

The background to this work is the notion of a trainable image analysis system; a scientist trains the system to find certain geological features by giving it examples of features to be located. The scientist can thus customize the tool to search for one type of feature versus another simply by providing positive and negative examples. In addition to automating laborious and visually-intensive tasks, the system provides an objective, examinable, and repeatable process for detecting and classifying objects in images. This allows scientists to base their analysis results on uniformly consistent data, free from subjective variations that invariably creep in when a visually exhausting task requiring many months or years is undertaken.

The Magellan spacecraft transmitted back to earth a data set consisting of over 30,000 high resolution synthetic aperture radar (SAR) images of the Venusian surface. This data set is greater than that gathered by all previous planetary missions combined — planetary scientists are literally swamped by data [Fayy94]. The study of volcanic processes is essential to an understanding of the geological evolution of the planet [Guest92]. Central to volcanic studies is the cataloging of each volcano location and its size and characteristics. There are estimated to be on the order of 10^6 visible volcanoes scattered throughout the 30,000 images [Aubele90]. Furthermore, it has been estimated that manually locating all of these volcanoes would require on the order of 10 man-years of a planetary geologist's time to carry out.

Empirical results using spatial eigenrepresentations (combined with supervised classification algorithms) have demonstrated that a trainable image analysis system can be roughly competitive with humans in terms of classification accuracy [Burl94, Fayy94]. The system uses a matched filter (for example, the mean of locally windowed training examples of volcanoes) to initially focus attention on local regions of interest. The detected local regions are projected into a subspace consisting of significant principal components of the training data. Although the full covariance matrix of the data (whose largest eigenvectors correspond to principal components) cannot be computed given the available sample sizes, the approximate eigenvectors can be computed using the singular value decomposition (SVD) technique [Burl94]. Supervised learning is used to produce a model which can discriminate between volcano and non-volcano local regions in the projected subspace. A simple maximum-likelihood multi-dimensional Gaussian classifier with full covariance matrices has been found to perform as well as alternative non-parametric methods such as neural networks and decision trees for the problem of discriminative learning in the projected eigenspace [Burl94].

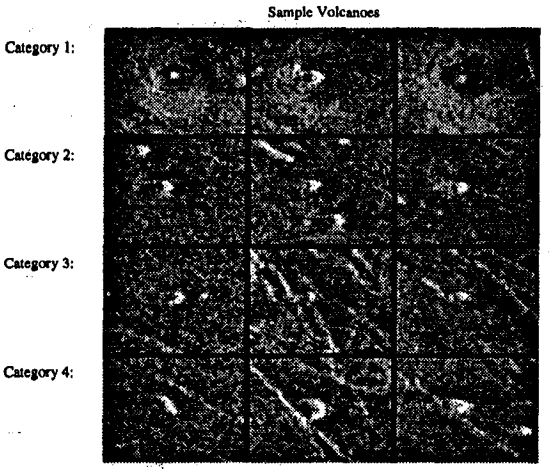


Figure 1: A small selection of volcanoes from four categories as labeled by the geologists.

3 Eliciting Ground Truth Estimates from Scientists

In the volcano location problem, as in many remote sensing applications, real ground truth data does not exist. No one has ever actually been to the surface of Venus (apart from a Russian robotic lander which melted within a few minutes), and despite the fact that the Magellan data is the best imagery ever obtained of Venus, scientists cannot always determine with 100% certainty whether a particular image feature is indeed a volcano.

In principle, for a given local region of interest, a scientist can provide a subjective probability that a volcano exists at that point *given* the local intensity values. It can been shown [Smyth94] that eliciting subjective probabilities is preferable to forcing a "yes/no" decision. In particular, for a fixed training sample size, probabilistic labels provide more information about the underlying Bayes-optimal discrimination boundary than "hard-decision" labels — hence, learning methods based on probabilistic labels will converge more quickly as the sample size increases. However, this result is conditioned on the assumption that the scientists are providing perfect unbiased subjective probability estimates. It is well known that accurate elicitation of subjective probabilities from humans is quite difficult and subject to various calibration errors and biases [Kahn82].

3.1 Defining Sub-Categories of Volcanoes

A more effective approach in practice is to have the scientists label training examples into quantized probability bins, where the probability bins correspond to visually distinguishable sub-categories of volcanoes. In particular, we have used 5 bins: (i) summit pits, bright-dark radar pair, and apparent topographic slope, all clearly visible, probability 0.98, (ii) only 2 of the 3 criteria in category (i) are visible, probability 0.80, (iii) no summit pit visible, evidence of flanks or circular outline, probability 0.60, (iv) only a summit pit visible, probability 0.50, (v) no volcano-like features visible, probability 0.0. The probabilities correspond to the mean probability that a volcano exists at a particular location given that it has been identified as belonging to a particular bin. These probabilities were elicited based on considerable discussions with the participating planetary geologists. How we use these probabilities for both training and evaluation will be discussed in more detail in the next few sections.

Figure 1 shows some typical volcanoes from each category. The use of quantized probability bins to attach levels of certainty to subjective image labels is not new: the same approach is routinely used in the evaluation of radiographic image displays to generate subjective ROC (receiver operating characteristic) curves [Chest92]. However, this paper extends the basic approach by defining the notion of probabilistic ROC curves (see Section 5).

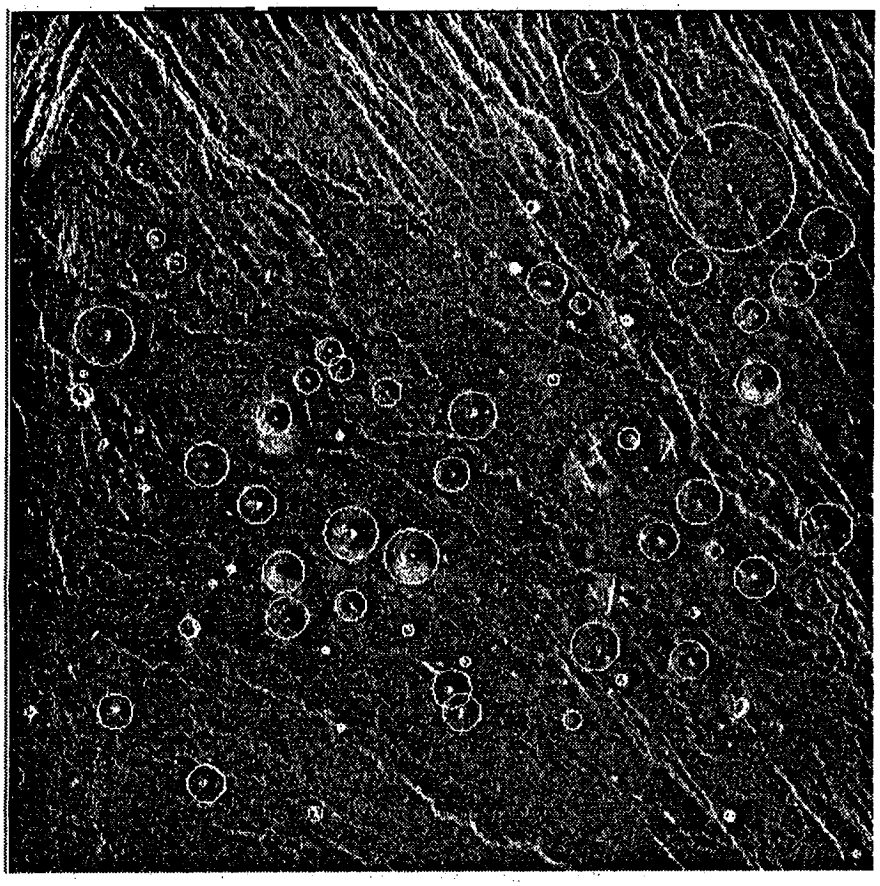


Figure 2: Magellan SAR image of Venus with consensus ground truth showing size and locations of small volcanoes.

3.2 Methodologies for Collecting Subjective Label Information

Participating in the development of the detection algorithm are planetary geologists from the Department of Geological Sciences at Brown University. We are fortunate to have direct collaboration with members of this group who have extensive experience in studying both earth-based and planetary volcanism and have published some of the standard reference works on Venus volcanism [Aubele90, Guest92]. Hence, their collective subjective opinion is (roughly speaking) about as expert as one can find given the available data and our current state of knowledge about the planet Venus.

It is an important point that, in the absence of absolute ground truth, the goal of our work is to be as comparable in performance as possible to the scientists in terms of labelling accuracy. Absolute accuracy is not measurable for this problem. Hence, the best the algorithm can do is to emulate the scientist's performance — this point will become clearer when we discuss performance metrics later in the paper.

A standard Magellan image consists of 1000×1000 pixels, where the pixels are 75m in resolution for the results referred to in this paper. Small volcano diameters are typically in the 2–3km range, i.e., 30 to 50 pixels wide. Volcanoes are often spatially clustered in volcano fields. As a consequence, most of the volcanoes are expected to be found in about 10-20% of the total number of images, and within these images there may number as many as 100 or more volcanoes, although typically the number is in the 10-50 range.

The standard manner in which we obtain labels is to have a labeller interact with a special X-windows interface implemented as part of the JARtool (JPL Adaptive Recognition Tool) software package tool being developed for use by scientists in analyzing the magellan data set. The user mouse-clicks locations of candidate volcanoes. Starting with a new image, the labeller proceeds to sequentially click on the estimated centers of the volcanoes and is then prompted to provide a subjective label estimate from a choice of categories 1–4 as described above. By default, locations which are not labeled are considered to have label "0" (non-

volcano). Clearly it is possible that based on the visual evidence, for the same local image patch, the same label may not be provided by different labellers, or indeed by the same labeller at different times. In addition to labels, the labeller can also provide a fitted diameter estimate by fitting a circle to the feature. Figure 2 shows a typical image labelled in this manner.

After completing the labelling, the result is an annotation of that image which can be stored in standard database format — the unique key to the image is a label event, which corresponds to a particular latitude/longitude (to the resolution of the pixels) for a particular labeller at a particular time (since the same labeller may relabel an image multiple times). It is this database which provides the basic reference framework for deriving estimates of geologic parameters, training data for the learning algorithms, and reference data for performance evaluation. A simple form of spatial clustering is used to determine which label events (from different labellers) actually correspond to the same geologic feature (volcano). It is fortunate that volcanoes tend not to overlap each other spatially and thus maintain a separation of at least a few kilometers, and also that different scientists tend to be quite consistent in their centering of the mouse-clicks — mean differences of about 2.5 pixels (Euclidean distance) have been found in cross comparisons of label data from different scientists, which is reasonable considering the precision one can expect from mouse location on a screen.

Table 1: Confusion Matrix of Scientist A Vs. Scientist B.

			Scientist A		
	Label 1	Label 2	Label 3	Label 4	Not Detected
Scientist B					
Label 1	19	8	4	1	3
Label 2	9	8	6	5	5
Label 3	13	12	18	1	37
Label 4	1	4	5	24	15
Not Detected	4	8	29	16	0

Table 1 shows the confusion matrix between two geologists for a set of 4 images. The (i, j)th element of the confusion matrix counts the number of label events which corresponded to labeller A generating label i and labeller B generating label j, where both labels were considered to belong to the same visual feature, i.e., were within a few pixels of each other. The $(i, 0)$ entries count the instances where labeller A provided label i, but labeller B did not provide any label — entry (0,0) is defined to be zero. Ideally, the confusion matrix would have all of its entries on the diagonal if both labellers agreed completely on all events. Clearly, however there is substantial disagreement, as judged by the number of off-diagonal counts in the matrix. For example, label 3's are particularly noisy, in both "directions." Label 3's are noisier than label 4's because there is less variability in the appearance of 4's compared to 3's (4's are simple pits, 3's are less well-defined). On the order of 50% of the label 3's detected by either labeller are not detected at all by the other labeller. On the other hand only on the order of 10% of the label 1's of either labeller are missed by the other. The matrix underlines the importance of modeling probabilistic labels for this particular problem.

4 Relating Probabilistic Labels to Ground Truth

The first step is to determine what it means when a scientist provides a particular label: what is the probability that a volcano actually exists given a label? We will use the shorthand v and $\bar{v}$ to denote the events "volcano present" and "volcano not present", respectively, and l to denote a particular label, $0 \leq l \leq l_{\max}$ ($l_{\max} = 4$ for the labelling problem). Let V be a binary random variable taking values v and $\bar{v}$, and let L be another discrete random variable taking values l, $1 \leq l \leq l_{\max}$. The shorthand notation of "v" for "$V = v$," etc., will be used. Note that we assume that labelling is *stochastic* rather than *deterministic* in the sense that presented multiple times with the same local image region, a scientist may not always provide the same label. The relevant probabilities we are interested in are conditional probabilities of the form $p(\text{volcano}|\text{label}) = p(v|l)$. Note that the difficulty in determining these probabilities lies in the fact that

the V is a hidden variable and cannot be observed directly.

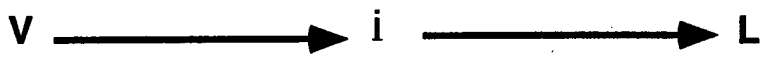


Figure 3: Causal Model 1 of Volcano Labelling Process.

Consider Figure 3 which identifies a simple causal model: volcanoes are mapped to an image intensity $\underline{i}$, which in turn is mapped to a label by the scientists. There is an implicit conditionalization on local pixel regions of fixed size, i.e., the labelling process is effectively treated as a series of decisions on such local regions. From Figure 3 we are ultimately interested in determining the probability of a volcano given $\underline{i}$. To train and evaluate our models we need to estimate terms such as $p(v|l)$. If we expand this out, we have to condition on all possible realizations of the image intensity $\underline{i}$;

$$p(v|l) = \sum_{\underline{i}} p(v|\underline{i}, l)p(\underline{i}|l) \tag{1}$$

Given the dimensionality of $\underline{i}$ (all possible intensities of a local region), this method of estimating $p(v|l)$ is clearly impractical. Note that the above equation can be rewritten as:

$$p(v|l) = \sum_{\underline{i}} p(v|\underline{i})p(\underline{i}|l) \tag{2}$$

since by the causal model of Figure 3, V is conditionally independent of L.

It is convenient to assume that the volcanoes correspond to visually distinguishable categories, namely "types." In addition, "type 0" will be used to identify all local images not covered by the "well-distinguished" types (i.e., non volcanoes in general). "Type" will be treated as another random variable T, taking values $1 \leq t \leq t_{\max}$, where $t_{\max} = l_{\max}$ typically. Conceptually it is useful to imagine that there is an Oracle who can unambiguously identify types given intensity information; the main point is that we do not have access to such an Oracle, but rather have access only to fallible scientists who provide labels, noisy estimates of types. In other words, T is an unobserved, hidden variable, while L is observed directly.

Figure 4: Causal Model 2 of Volcano Labelling Process: volcanoes to types to labels.

To circumvent the problems of estimating probabilities conditioned on intensity values the following simplification of the model is proposed: replace the high dimensional intensity $\underline{i}$ with the low-dimensional T in the causal model. T can be considered a quantization of the intensity map. The effect is to remove any dependency on intensity values in the model, which can now be written as

$$p(v|l) = \sum_{t} p(v|t)p(t|l) \tag{3}$$

The dependence of types on volcanoes will be assumed given by the scientists as a general piece of prior information — in particular, $p(v|t)$, for $t = 1, 2, 3, 4$ are the subjective probabilities we have elicited from the scientists which described the mean probability that a volcano exists at a particular location, given that it belongs to a particular type (Section 3.1). These subjective probabilities are not conditioned on *labels* per se, but on the *types*, i.e., $p(v|t) \in \{0.98, 0.80, 0.60, 0.5, 0.0\}, t \in \{1, 2, 3, 4, 0\}$.

The $p(t|l)$ terms in Equation (3) represent the estimation noise resulting from the fact that scientists are unable to specify, with 100% certainty, the particular "type" of a volcano. Determination of these probabilities is rendered non-trivial by the fact that the true types t are never directly observed, and thus some assumptions about the relationship between T and L must be made in order to infer their dependence. At this point in time, estimating the $p(t|l)$ terms from multiple labellings of the same data represents work in progress — a particular method is outlined in Appendix 1. Note that the overall effect of the above models

will be to reduce our confidence that a typical local region is a volcano, given some labelling information — this has direct implications for estimating the overall numbers of volcanoes in a particular region, and so forth. For example local regions which which produce label disagreements between experts will be down-weighted compared to volcanoes which receive unanimous labels.

5 Performance Evaluation: Probabilistic Free-Response ROC Analysis

Given that the scientists cannot classify each object with 100% confidence, how can we assess how well our algorithms are performing? We have investigated the idea of "consensus ground truth": a consensus-based probabilistic labelling is generated by multiple scientists working together on labelling images. The individual labellings and the results of the automated detection system described earlier are then evaluated in terms of performance relative to the consensus. The performance of an algorithm is considered to be satisfactory if, compared to consensus ground truth, its performance is as good as that of an individual scientist. There is an implicit assumption that the consensus process is "fair" in the sense that the results reflect the measured expertise of each member and is not biased by factors such as personality traits, seniority, and so forth. For the volcano application we believe that the consensus is reasonably fair in this sense.

As a performance evaluation tool we use a variation of the well-known receiver operator characteristic (ROC) methodology. The purpose of the ROC is to determine the complete range of performance of a decision system in terms of its estimated detection rate versus false alarm rate. Consider a binary hypothesis testing problem (equivalently a binary classification or discrimination problem): the 2 mutually exclusive and exhaustive hypotheses are denoted as ω_1 and ω_2. Let $\underline{x}$ denote the observed data.

Standard Bayesian decision theory [VanTrees68] shows that the optimal decision rule (in terms of minimum cost) must be of the form:

$$\textbf{If } \frac{p(\omega_1|\underline{x})}{p(\omega_2|\underline{x})} \geq t \textbf{ then choose } \omega_1, \textbf{ else choose } \omega_2 \tag{4}$$

where t, $0 \leq t \leq \infty$, is a particular decision threshold (related to the various types of misclassification costs). Hence, as t is varied one obtains different decision rules. Consider what happens whenever the rule chooses ω_1. If the true state of nature is actually ω_1 then a *detection* occurs, whereas if the true state of nature is ω_2 then a false alarm occurs: the probabilities of detection, $p_d(t)$, and false alarm, $p_{fa}(t)$, are thus defined as the respective probabilities that a detection or false alarm occurs given that the rule chooses ω_1. As $t \rightarrow \infty$ the rule becomes more conservative: fewer detections, but also fewer false alarms. As $t \rightarrow 0$ the rule becomes more liberal: more detections, but also more false alarms.

When the conditional densities $p(\omega_1|\underline{x})$ and $p(\omega_2|\underline{x})$ are known exactly one can determine $p_d(t)$ as a function of $p_{fa}(t)$; this plot is known as the ROC and provides the characteristic signature of a decision system over the entire range of possible detection/false alarm operating points. The utility of this framework is that one can evaluate the performance of a particular decision rule over a range of possible operating values for t and thus determine a useful operating point, e.g., fix the false alarm rate at (say) 20%.

Since in practical applications $p(\omega_1|\underline{x})$ and $p(\omega_2|\underline{x})$ are unknown, the ROC must be estimated directly from data. This is straightforward provided the decision system is producing either a direct estimate of the ratio $r = \frac{p(\omega_1|\underline{x})}{p(\omega_2|\underline{x})}$, or some monotonic function of r. The estimation procedure is to vary r (or a monotonic function of same) as a decision threshold on a labeled training data set and count the resultant numbers of detections and false alarms for each value of r. A training set of size N produces $N + 1$ operating points (including the end points of (0.0,0.0) and (1.0, 1.0)). One converts the number of detections and number of false alarms to probabilities by dividing by the total number of training examples of class ω_1 and class ω_2 respectively. Thus, one can plot an empirical ROC, the estimated probability of detection versus estimated probability of false alarm.

For the volcano detection problem, the reference labels are taken from the consensus labelling, i.e., this is in effect treated as ground truth. Class ω_1 corresponds to volcanoes, class ω_2 to non-volcanoes. False alarms correspond to label events which are categorized by the detection system as being of class volcano, when the consensus labelling indicates a non-volcano event, i.e., a local region which was not labeled. There

is a problem in defining the probability of false alarm, since it is difficult to define the prior probability of class ω_2. For example, should the prior probability be proportional to the number of pixels in the image which were not labeled as volcanoes ? This definition does not make much intuitive sense, since it would be a function of the number of pixels in a given image (one wants the ROC to be invariant to changes in such parameters) and also since it would result in an astronomically high prior in favour of non-volcanoes.

Hence, we use detection rate versus *false alarms per total number of detections* — this normalized false alarm rate is a much more useful parameter since it is invariant to the size and resolution of the images used to determine the ROC. This plot is no longer directly interpretable as a standard ROC since the false alarm rate axis can now run from 0% to some arbitrary percentage greater than 100%, i.e., there may have been more false alarms detected than true detections in total for some threshold operating points. This slightly modified ROC methodology is essentially the same as the free-response ROC (FROC) used in evaluation of radiology display systems [Bunch78, Chakra90].

Furthermore, standard ROC and FROC approaches assume that ground truth is known. When ground truth is known only in probabilistic form as described earlier, one must allow for the fact that each detected local region is only a detection with probability p: there is an associated probability of $1 - p$ of it being a false alarm. These probabilities are determined with reference to the consensus labelling: for example, for a given threshold value, if the detection system being evaluated detects a particular local region which has been labeled by the consensus as category 2 (probability of volcano = 0.8), then this is counted as 0.8 of a detection and 0.2 of a false alarm. The overall effect is to drag the non-probabilistic ROC (where no allowance is made to for the probabilistic effect) towards the "center" of the plot, away from the ideal "false alarm rate 0.0, detection rate 1.0" operating point. Furthermore, the ideal "perfect" operating point is no longer achievable by *any* system, since the reference data is itself probabilistic. Hence, an effective optimal ROC is defined by exactly matching the probabilistic predictions of the consensus — one can do no better. We denote the probabilistic method with the normalized false alarm rate as the probabilistic FROC (PFROC) (in [Henk90] a similar concept to the PFROC is defined for the special case of multivariate normal distributions).

Within this framework, the performance of a human labeller can only be determined within the resolution of the quantized probabilistic bins used in the subjective labelling process. With k bins, one can determine the location of k operating points on the PFROC, including the (0.0, 0.0) point.

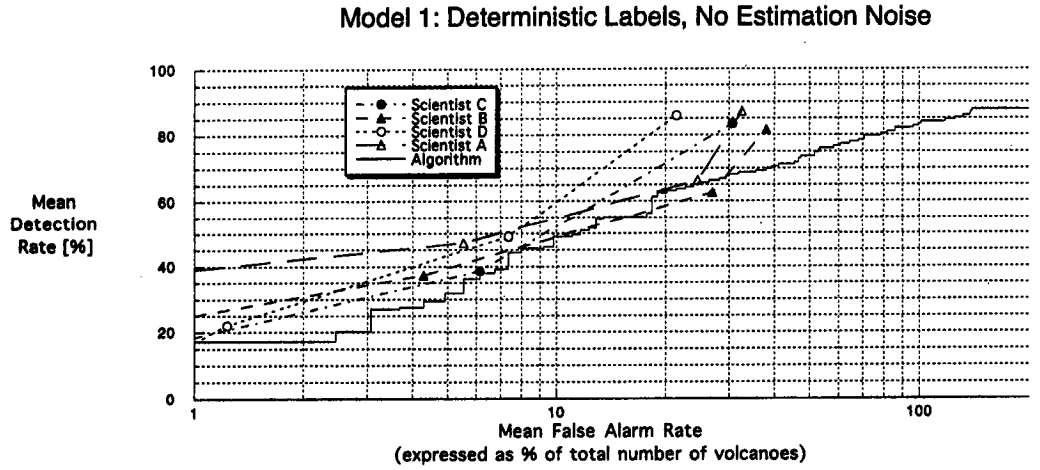
Figure 5 shows a PFROC curve for four images, comparing the performance of 4 planetary geologists and the current version of the detection algorithm. A consensus of 2 planetary geologists was used as reference. The consensus labelling was determined some time *after* the individual labellings from the same scientists. The algorithm was as described in Section 2 (and in more detail in [Burl94]) and was evaluated in cross-validation mode (trained on 3 of the images, and tested on the fouth, repeated four times). In total, the consensus labelling produced 163 volcanoes, which correspond to the 100% point on the y-axis: as described above, the false alarm rates are determined relative to the 163 "true" detections.

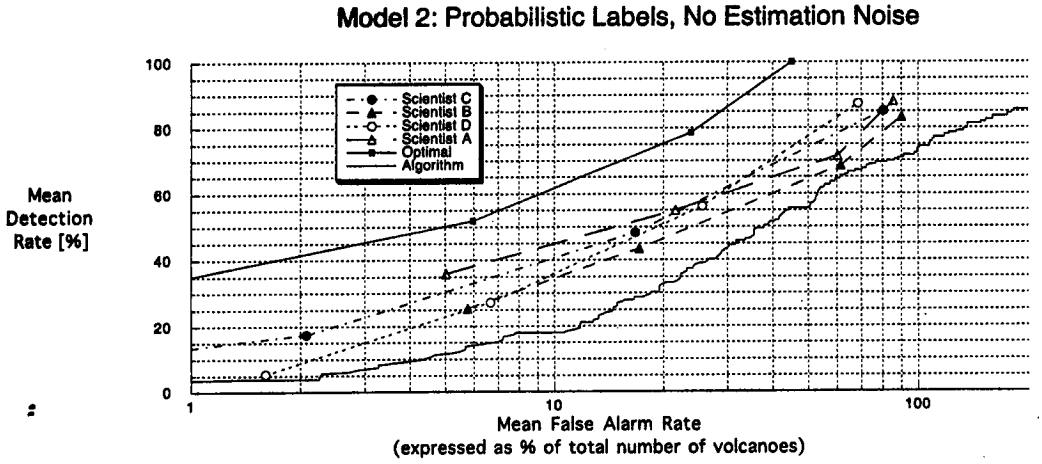
The top curve is that which is obtained if one ignored the probabilistic nature of the labels completely and just treated all samples with labels {1,2,3,4} as volcanoes. The scientist's upper operating points are at about 80% detection and 30% false alarm rates, with the algorithm begin competitive in performance for the 5–20% range of false alarms.

The other two curves were determined using the PFROC described above: the center plot corresponds to assuming no estimation noise ($p(t|l) = 1.0, t = l$), whereas the lower plot contains an estimation noise model. The estimation noise probability matrix was estimated as described in the Appendix. For each curve, the relative weightings assigned to each detection or false alarm were equal to the estimated $p(v|l)$, where l is the label attached to each test example according to the consensus labelling.

The upper curve in each plot ("Optimal") is the upper bound reflecting how well a scientist or an algorithm could do if they matched the consensus labelling exactly. The other curves largely parallel this curve but are 10 to 50% less accurate in terms of detection rate at a fixed false alarm rate. The effect of introducing probabilistic labels is to induce a dramatic shift (to the right) in terms of increased false alarm rate, i.e., compared to the top plot, the scientist's upper operating point is now at 80-90% false alarms for about an 80% detection rate. This is a direct consequence of the fact that many of the volcanoes which are being detected only have a 60% chance of actually being a volcano.

Furthermore, the curve for the algorithm is now further away from those of the scientists. This can be explained by the fact that the detection algorithm has a relatively poor ability to accurately determine





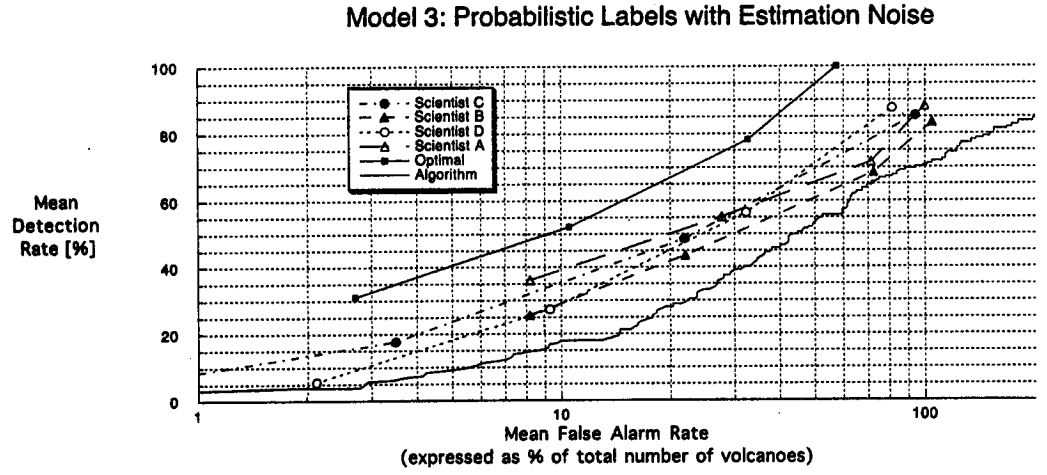


Figure 5: (i) standard FROC, (ii) probabilistic FROC (PFROC) with no estimation noise, (iii) PFROC with estimation noise

posterior class probabilities, in that a class 3 volcano might be estimated to have posterior probability 0.9, whereas a class 1 volcano might only be given a probability of 0.6. This distinction is penalized by the probabilistic ROC method, whereas it is not penalized by the conventional approach (top plot). Thus, the PFROC provides more information to the algorithm designer in terms of evaluating how well the algorithm performance matches that of the scientists. We note that the probabilistic nature of the labels (the center plot) has the major effect: adding estimation noise (lower plot) only causes a slight deterioration in estimated performance.

The PFROC clearly demonstrates that both the scientists and the algorithm are relatively inaccurate at detecting volcanoes. Nevertheless, we feel these results provide a true picture of the relative accuracy with which volcanoes can be detected in the Magellan images. This underlying ambiguity in volcano detectability should be recognized and factored into any scientific inferences made based upon labellings by individuals or machine algorithms.

6 Other Aspects of Probabilistic Labels

The acknowledgment of uncertainty in the labelling can have other significant impacts on overall image analysis methodologies. For example, as described in detail in [Smyth94] and [Burl94], the matched filter generation, SVD subspace generation, and discriminant learning procedures can all be modified to account for probabilistic labels. The general approach is based on the notion of assigning fractions of a training data sample to each class in proportion to the subjective label weight: for example, a category 2 volcano might be treated as 0.8 of a sample for the volcano class and 0.2 of a sample for the non-volcano class. When the estimation noise can be calibrated and there are labels from multiple experts, the methods of the Appendix can be used to determine more accurate relative class weighting. While this weighted treatment of probabilistic labels leads to improved performance in theory [Smyth94], in our experiments to date we have found no improvement in performance by learning from probabilistic labels as compared to the default approach of treating all labelled items as examples of class volcano. Investigation of the data revealed that the subspace projection technique was destroying any probabilistic structure which existed in the data at the level of the intensity maps, i.e., category 1's, 2's, 3's and 4's were all being projected into the same region of feature space (as revealed by 2-d scatterplots of various feature pairs) and completely overlapped each other without any structure. If the probabilistic structure had been preserved, one would expect to see the 1's to be further away from the non-volcano class than the 2's and so forth. This is an example of a learning algorithm dealing with a feature space (SVD filter responses) which is different than that on which the labelling is performed (local intensity maps), with the result that the probabilistic labels do not relate in any useful way to the space in which learning is taking place. As a consequence, a detection algorithm based only on SVD filter responses cannot reproduce accurate posterior probability estimates which match those of the scientists subjective labels. A current direction of investigation is to seek projections which preserve the probabilistic label information, which in turn should result in better PFROC performance.

Estimation of various spatial statistics can also be conditioned on the probabilistic nature of the labels — for example, non-parametric kernel density estimates of the volcano diameters (an important geological "signature") can be modified to take probabilistic labels into account as described in [Smyth94]. Densities which are not unimodal are particularly sensitive to probabilistic labels: incorrect treatment of the labels can lead to oversmoothing of real modes, or the introduction of spurious ones. Once again, the actual values of the probabilistic labels are a function of the particular noise model one chooses to use as described in the Appendix. Estimation of spatial statistics in this manner is a topic of current investigation.

7 Conclusion

The major focus of this paper is the treatment of uncertainty in the training data when designing and evaluating knowledge discovery systems for image databases. The net effect of ground truth ambiguity is to propagate an extra level of subjective noise into processes such as training learning algorithms, performance evaluation methodologies, and estimation of spatial statistics of science interest. Handling this uncertainty requires the introduction of special techniques such as the probabilistic free-response ROC (PFROC) methodology. The proposed techniques provide a framework for more accurate estimation and evaluation of basic

image quantities of interest for applications where absolute ground truth is not available. Such applications are becoming increasingly common as remote-sensing platforms provide orders of magnitude more data and well-calibrated ground truth constitutes a tiny (and perhaps even zero) fraction of the overall data set.

The Magellan volcano detection problem provides a clear example of this problem in a realistic and large-scale setting. The problem of lack of ground truth is fairly pervasive and is applicable to a wide range of domains covering remote sensing, medical diagnosis, and large corporate and financial database analysis. In a typical large scale knowledge discovery problem, uncertainties in training data (or examples of patterns of interest) need to be treated and approximated explicitly. Ignoring this information, as is often done, by assuming class labels are absolute is likely to result in a miscalibrated system. This paper also discusses issues that arise in measuring the performance of the learning system in this context.

Acknowledgments

The research described in this report has been carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. We would like to thank our collaborating geologists: Jayne Aubele and Larry Crumpler of Brown University for their assistance in labelling and analyzing the Magellan data. We thank Maureen Burl, John Loch, Jennifer Yu, and Joe Roden for work on developing the software and user-interfaces of JARtool. This work is sponsored by by NASA OACT - Code CD.

References

[Agresti92] Agresti, A., (1992), "Modelling patterns of agreement and disagreement," *Statistical Methods in Medical Research*, vol.1, pp.201–218.

[Aubele90] Aubele, J. C. and Slyuta, E. N. (1990), "Small domes on Venus: characteristics and origins," in *Earth, Moon and Planets*, **50/51**, 493–532.

[Chakra90] Chakraborty, D. P., and Winter, L. H. L. (1990), " Free-Response methodology: alternate analysis and a new observer-performance experiment," *Radiology*, 174, 873–881.

[Bunch78] Bunch, P. C., Hamilton, J. F., Sanderson. G. K., and Simmons, A. H., (1978), "A Free-Response approach to the measurement and characterization of radiographic-observer performance," *J. Appl. Photo. Eng.*, vol.4, no. 4, pp.166-171.

[Burl94] Burl, M.C., Fayyad, U.M., Perona, P., Smyth, P., and Burl, M.P. (1994), "Automating the hunt for volcanoes on Venus," in *CVPR-94: Proceedings of the 1994 Computer Vision and Pattern Recognition Conference*, to appear.

[Chest92] Chesters, M. S., (1992), "Human visual perception and ROC methodology in medical imaging," *Phys. Med. Biol.*, vol.37, no.7, pp.1433-1476.

[Fayy93] Fayyad, U. M., and Smyth, P., (1993) "Image database exploration: progress and challenges," *Proceedings of the 1993 AAAI Workshop on Knowledge Discovery in Databases.*

[Fayy94] Fayyad, U. M., P. Smyth, N. Weir, and S. Djorgovski (1994), "Automated analysis and exploration of large image databases: results, progress, and challenges, " *Journal of Intelligent Information Systems*, in press.

[Guest92] Guest, J. E. et al. (1992). "Small volcanic edifices and volcanism in the plains of Venus," *Journal of Geophysical Research*, vol.97, no.E10, pp.15949-66.

[Henk90] Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990) "Receiver operator characteristic (ROC) analysis without ground truth," *Medical Decision Making*, vol.10, no.1, pp.24–29.

[Kahn82] Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982), *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

[Lug92] Lugosi, G., (1992) "Learning with an unreliable teacher," *Pattern Recognition*, vol. 25, no.1, pp.79–87.

[Science91] *Science, special issue on Magellan data*, April 12, 1991.

[Silver80] Silverman, B., (1980), "Some asymptotic properties of the probabilistic teacher," *IEEE Trans. Info. Theory*, IT-26, no.2, pp.246–249.

[Smyth94] Smyth, P., (1994), "Learning with probabilistic supervision," in *Computational Learning Theory and Natural Learning Systems 3*, T. Petsche, M. Kearns, S. Hanson, R. Rivest (eds), Cambridge, MA: MIT Press, to appear.

[Ueber93] Uebersax, J. S., (1993), "Statistical modeling of expert ratings on medical treatment appropriateness," *J. Amer. Statist. Assoc.*, vol.88, no.422, pp.421–427.

[VanTrees68] Van Trees, H. L., (1968), *Detection, Estimation, and Modulation Theory: Part 1*, John Wiley and Sons, New York, pp.23–31.

Appendix: Estimating the $p(t|l)$ terms from multiple labellings

Consider that we have a database of N labelled local regions. Assume that each local region has been examined m times, either by m different scientists or groups of same, or the same scientist multiple times, or some combination of same (the extension to the case where some subsets of local regions have been examined by different numbers of labellers or groups of labellers is trivial and will be omitted to keep notation simple). Hence, each local region has been labelled as one of the 4 labels 1/2/3/4 by at least one of the m labellers.

For each label event assign a "vote" of $1/m$ to each label 1/2/3/4 each time that a labeller assigns that label, and a "vote" of $1/m$ to the label 0 if a labeller did not label it at all. Implicit here is an assumption that each labeller is being weighted equally — extensions to the case of non-equal weighting are straightforward and are not dealt with here. We can interpret the sum of the votes for a particular label (from different labellers) as the probability that local intensity $\underline{i}$ will be assigned label l. More formally, we define the estimator

$$\hat{p}(l|\underline{i}) = \frac{1}{m}\sum_{k=1}^{m}\delta(l, v_k(\underline{i})) \tag{5}$$

where $\delta(x,y) = 0$ unless $x = y$, and $v_k(\underline{i})$ is the label provided by the kth labeller for the local intensity map $\underline{i}$.

We can now estimate the marginal probability that an arbitrary labeller will assign label l to a local region, by summing over all intensities:

$$\hat{p}(l) = \sum_{j=1}^{N}\hat{p}(l|\underline{i})\hat{p}(\underline{i}^j) = \frac{1}{N}\sum_{j=1}^{N}\hat{p}(l|\underline{i}^j) \tag{6}$$

where j is an index over the N local regions in the database. To estimate $p(t|l)$ by Bayes' rule we first need to estimate $p(t,l)$. The following estimator is defined:

$$\begin{aligned}\hat{p}(t,l) &= \sum_{\underline{i}}\hat{p}(t,\underline{i},l) = \sum_{\underline{i}}\hat{p}(t|l,\underline{i})\hat{p}(l|\underline{i})p(\underline{i}) = \frac{1}{N}\sum_{j=1}^{N}\hat{p}(t|l,\underline{i}^j)\hat{p}(l|\underline{i}^j)\\ &= \frac{1}{N}\sum_{j=1}^{N}\hat{p}(t|\underline{i}^j)\hat{p}(l|\underline{i}^j)\end{aligned}$$

since the type t is independent of the label l given the local intensity $\underline{i}$. If we define the estimator for $\hat{p}(t|\underline{i}^j)$ to be the same as for $\hat{p}(l|\underline{i}^j)$ (as in Equation 6 above), the process is complete, since all necessary terms are now defined and can be estimated directly from the database. Finally, we have that

$$\hat{p}(t|l) = \frac{\hat{p}(t,l)}{\hat{p}(l)} \tag{7}$$

As an example, this method was applied to two labellings of the same 9 images with the following results:

$$\begin{aligned}\hat{p}(T=1|L=1) &= 0.80, \quad \hat{p}(L=1) = 0.11\\ \hat{p}(T=2|L=2) &= 0.68, \quad \hat{p}(L=2) = 0.22\\ \hat{p}(T=3|L=3) &= 0.70, \quad \hat{p}(L=3) = 0.29\end{aligned}$$

$$\begin{aligned}\hat{p}(T=4|L=4) &= 0.78, \quad \hat{p}(L=4) = 0.22\\ \hat{p}(T=0|L=0) &= 0.5, \quad \hat{p}(L=0) = 0.16\end{aligned}$$

Labelling of 1's and 4's appears to be the most accurate, labellings of 2's and 3's less so. Furthermore, it is estimated that 16% of the local regions identified (out of 330 which were labelled by at least one labeller in the 9 images) are truly non-volcanoes.

The replacement of $\hat{p}(t|\underline{i}^j)$ by $\hat{p}(l|\underline{i}^j)$ above is a provably biased approximation but the method appears to yield reasonable results in practice. The properties of this estimator are currently under investigation as are alternative estimators, e.g., methods based on maximum likelihood estimation techniques using parametric models of hidden (latent) traits [Ueber93, Agresti92].

In addition it should be noted that the method described above implicitly assumes that all experts are treated equally in the sense that a single average estimation matrix is derived for all. A more sophisticated approach would be to produce estimation noise models for each individual expert. Although there are some subtleties to this aspect of the problem, in principle it is solvable by the same techniques as described above.