Towards Exploratory Test Instance Specific Algorithms for High Dimensional Classification

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
charu@us.ibm.com

ABSTRACT

In an interactive classification application, a user may find it more valuable to develop a diagnostic decision support method which can reveal significant classification behavior of exemplar records. Such an approach has the additional advantage of being able to optimize the decision process for the individual record in order to design more effective classification methods. In this paper, we propose the Subspace Decision Path method which provides the user with the ability to interactively explore a small number of nodes of a hierarchical decision process so that the most significant classification characteristics for a given test instance are revealed. In addition, the SD-Path method can provide enormous interpretability by constructing views of the data in which the different classes are clearly separated out. Even in cases where the classification behavior of the test instance is ambiguous, the SD-Path method provides a diagnostic understanding of the characteristics which result in this ambiguity. Therefore, this method combines the abilities of the human and the computer in creating an effective diagnostic tool for instance-centered high dimensional classification.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms

Keywords

visual data mining, classification

1. INTRODUCTION

High dimensional data is a challenge to subspace based classification methods such as the decision tree because of the large number of combinations of dimensions (or subspaces) which have classification power. The basic limitation of such methods is that they try to create a succint summary of small number of discriminatory subspaces from an exponential number of possibilities. The particular pattern which is most suitable for the classification of a given record is *specific to that record* and could be present in one of many partially overlapping subspace clusters. The succint summary may fail to capture such instance-specific characteristics. This incompleteness in data characterization may result in the particular structure of the tree to be more or less suited to particular kinds of test instances. We note that this incompleteness problem extends to most classification models such as rule based systems, neural networks, or bayesian methods in which the aim is to create a *summarized and efficiently usable model* of the relationship between the feature and class variables [3], rather than providing comprehensive exploratory ability for individual test instances.

In this paper, we propose an open-ended and *test-instance specific* hierarchical decision *process* in which the primary aim is to provide diagnostic ability. We note that this scheme is *not* intended as an alternative to current batch-processing methods, which are relevant to classifying large numbers of test instances. This approach is more suitable for cases where detailed diagnostic information is required about individual test instances during the classification process. We note that most previous interactive approaches [2] create a decision tree on the entire data set, and is therefore not suited to test-instance specific diagnosis.

This paper is organized as follows. The quantification of instance-specific subspaces is discussed in section 2. In section 3, we will utilize this quantification to develop the subspace decision path method. The empirical results are presented in section 4, and the conclusions are discussed in section 5.

2. SUBSPACE DISCRIMINATION

In order to identify an appropriate quantification of the level of separability of the different classes in subspaces, we need a technique which is particularly sensitive to the class discrimination behavior in the locality of a given point, and is also amenable to the determination of arbitrarily shaped patterns. One way of intuitively characterizing the discrimination in a subspace is to quantify the difference in class distribution at each point in the space. To this effect, we use the method of kernel density estimation.

In kernel density estimation, we find a continuous estimate of the density of a set of points. Let us assume that we have

KDD'05, August 21–24,2005,Chicago,Illinois,USA.

a data set $\mathcal{D}$ with N points and dimensionality d. The set of points in $\mathcal{D}$ are denoted by $X_1 \ldots X_N$. Let us further assume that the k classes in the data are denoted by $\mathcal{C}_1 \ldots \mathcal{C}_k$. The number of points belonging to the class $\mathcal{C}_i$ is n_i, so that $\sum_{i=1}^{k} n_i = N$. We assume that the data set associated with the class i is denoted by $\mathcal{D}_i$. The probability density at a given point is determined by the sum of the smoothed values of the kernel functions $K_h(\cdot)$ associated with each point in the data set. Thus, the density estimate of the data set $\mathcal{D}$ at the point x is defined as follows:

$$f(x, \mathcal{D}) = (1/n) \cdot \sum_{X_i \in \mathcal{D}} K_h(x - X_i) \quad (1)$$

The kernel function is a smooth unimodal distribution such as the gaussian function:

$$K_h(x - X_i) = 1/(\sqrt{2\pi} \cdot h) \cdot e^{-\frac{||x - X_i||^2}{2h^2}} \quad (2)$$

We note that the kernel function is dependent on the use of a parameter h which is the level of smoothing. The accuracy of the density estimate depends upon this width h, and several heuristic rules have been proposed for estimating the bandwidth. The widely used Silverman rule [4] sets $h = 1.06 \cdot \sigma \cdot N^{-1/5}$, where σ^2 is the variance of the N data points.

We note that the value of the density $f(x, \mathcal{D})$ may differ considerably from $f(x, \mathcal{D}_i)$ because of the difference in distributions of the different classes. Correspondingly, we define the *accuracy density* $\mathcal{A}(x, \mathcal{C}_i, \mathcal{D})$ for the class $\mathcal{C}_i$ as follows:

$$\mathcal{A}(x, \mathcal{C}_i, \mathcal{D}) = n_i \cdot f(x, \mathcal{D}_i)/(\sum_{i=1}^{k} n_i \cdot f(x, \mathcal{D}_i)) \quad (3)$$

We note that the above expression always lies between 0 and 1. The higher this value, the greater the relative density of $\mathcal{C}_i$ compared to the other classes. We further note that the sum of the accuracy values over the different classes is equal to one.

$$\sum_{i=1}^{k} \mathcal{A}(x, \mathcal{C}_i, \mathcal{D}) = 1 \quad (4)$$

Another related measure is the *interest density* at a given point x. The interest density of the class $\mathcal{C}_i$ at x is denoted by $\mathcal{I}(x, \mathcal{C}_i, \mathcal{D})$, and is defined in an analogous way to the accuracy density. In this case, the interest density is the ratio of the density of the class $\mathcal{C}_i$ to the overall density of the data. Therefore, we define the *interest density* $\mathcal{I}(x, \mathcal{C}_i, \mathcal{D})$ for the class $\mathcal{C}_i$ as follows:

$$\mathcal{I}(x, \mathcal{C}_i, \mathcal{D}) = f(x, \mathcal{D}_i)/f(x, \mathcal{D}) \quad (5)$$

The class $\mathcal{C}_i$ is over-represented at x, when the interest density is larger than one. The interest density is more revealing from a statistical perspective than the accuracy density, since it is not biased by the initial distribution of classes. The *dominant class* at the coordinate x is denoted by $\mathcal{CM}(x, \mathcal{D})$, and is equal to $\text{argmax}_{i \in \{1, \ldots k\}} \mathcal{I}(x, \mathcal{C}_i, \mathcal{D})$. Correspondingly, the maximum interest density at x is denoted by $\mathcal{IM}(x, \mathcal{D}) = \max_{i \in \{1, \ldots k\}} \mathcal{I}(x, \mathcal{C}_i, \mathcal{D})$. Both the interest and accuracy density are valuable quantifications of the level of dominance of the different classes. The interest density is more effective at comparing among the different classes at a given point, whereas the accuracy density is more effective at providing an idea of the absolute accuracy at a given point.

So far, we have assumed that all of the above computations are performed in the full dimensional space. However, we can also project the data onto the subspace E in order to perform this computation. Such a calculation would quantify the discriminatory power of the subspace E at x. In order to denote the use of the subspace E in any computation, we will superscript the corresponding expression with E. Thus the density in a given subspace E is denoted by $f^E(\cdot, \cdot)$, the accuracy density by $\mathcal{A}^E(\cdot, \cdot, \cdot)$, and the interest density by $\mathcal{I}^E(\cdot, \cdot, \cdot)$. Similarly, the dominant class is defined using the subspace-specific interest density at that point, and the accuracy density profile is defined for that particular subspace. An example of the accuracy density profile (of the dominant class) in a 2-dimensional subspace is illustrated in Figure 2(a). The test instance is also labeled in the same figure in order to illustrate the relationship between the density profile and test instance. We note that for large data sets estimated using the kernel density technique, the interest density profiles would have exactly the same shape as the accuracy density profiles, except that they are scaled differently.

The subspace specific determination of the interest density $\mathcal{I}^E(t, \mathcal{C}_i, \mathcal{D})$ at the test instance t is quite valuable, since it can be used to determine those characteristics which are most revealing about the class behavior of t. In order to do so, one may wish to find those projections of the data in which the interest density value $\mathcal{IM}^E(t, \mathcal{D})$ is the maximum. It is quite possible that in some cases, different subspaces may provide different information about the class behavior of the data; these are the difficult cases in which a test instance may be difficult to classify accurately. In such cases, the user may need to isolate particular data localities in which the class distribution is further examined by a hierarchical exploratory process.

3. THE EXPLORATORY PROCESS

In this section, we discuss methods for exploratory construction of decision paths. For a given test example, the end user is provided with unique options in exploring various characteristics which are indicative of its classification. To this effect, we use the subspace determination process discussed in the previous section. The subspace determination process finds the appropriate *local* discriminative subspaces for a given test example. These are the various possibilities (or branches) of the decision path which can be utilized in order to explore the regions in the locality of the test instance. In each of these subspaces, the user is provided with a visual profile of the accuracy density. This profile provides the user with an idea of which branch is likely to lead to a region of high accuracy for that test instance. This visual profile can also be utilized in order to determine which of the various branches are most suitable for further exploration. Once such a branch has been chosen, the user has the option to further explore into a particular region of the data which has high accuracy density. This process of data localization can quickly isolate an arbitrarily shaped region in the data containing the test instance. This sequence of data localizations creates a path (and a locally discriminatory combination of dimensions) which reveals the underlying classification causality to the user.

In the event that a decision path is chosen which is not

Algorithm *SubspaceDecisionPath(Test Inst.: t, Data: $\mathcal{D}$, MaxDim: l, MaxBranchFactor: b_{max}, MinIRatio: ir_{min})*
begin
  PATH= $\{\mathcal{D}\}$;
  while *not(termination)* **do**
  begin
    Pick the last node $\mathcal{L}$ indicated in PATH;
    $\mathcal{E} = \{E_1 \ldots E_q\} =$
      ComputeClassifSubspaces($\mathcal{L}$, t, l, b_{max}, ir_{min});
    for each E_i *ConstructDensityProfile*(E_i, $\mathcal{L}$, t);
    if (*zoom-in (user-specified)*) **then**
    begin
      User specifies choice of branch E_i;
      User specifies accuracy den. thresh. λ for zoom-in;
      { $p(\mathcal{L}', \mathcal{C}_i)$ is accuracy significance of class
        $\mathcal{C}_i$ in $\mathcal{L}'$ with respect to $\mathcal{L}$ }
      $(\mathcal{L}', p(\mathcal{L}', \mathcal{C}_1) \ldots p(\mathcal{L}', \mathcal{C}_k)) = IsolateData(\mathcal{L}, t, \lambda)$;
      Add $\mathcal{L}'$ to the end of PATH;
    end;
    else begin (retreat))
      User specifies data set $\mathcal{L}'$ on PATH to backtrack to;
      Delete all data pointers occuring after $\mathcal{L}'$ on PATH;
    end;
    { Calculate cum. dominance of each class $\mathcal{C}_i$ along
      PATH in order to provide the user a measure of its
      significance }
    for each class $\mathcal{C}_i$ **do**
     $CD(PATH, \mathcal{C}_i) = 1 - \pi_{(\mathcal{L} \in PATH, \mathcal{L} \neq \mathcal{D})}(1 - p(\mathcal{L}, \mathcal{C}_i))$;
     Output $CD(PATH, \mathcal{C}_i)$;
  end;
end

Figure 1: The Subspace Decision Path Method

strongly indicative of any class, the user has the option to backtrack to a higher level node and explore a different path of the tree. In some cases, different branches may be indicative of the test example belonging to different classes. These are the "ambiguous cases" in which a test example could share characteristics from multiple classes. Many standard modeling methods may classify such an example incorrectly, though the subspace decision path method is much more effective at providing the user with an intensional knowledge of the test example because of its exploratory approach. This can be used in order to understand the causality behind the ambiguous classification behavior of that instance.

The overall algorithm for decision path construction is illustrated in Figure 1. The input to the system is the data set $\mathcal{D}$, the test instance t for which one wishes to find the diagnostic characteristics, a maximum branch factor b_{max}, and a minimum interest density ir_{min}. In addition, we input the maximum dimensionality l of any subspace utilized in data exploration. The value of $l = 2$ is especially interesting because it allows for the use of visual profile of the accuracy density. We note that even though it is natural to use 2-dimensional projections because of their visual interpretability, the data exploration process along a given path reveals a higher dimensional combination of dimensions which is most suitable for the test instance. The branch factor b_{max} is the maximum number of possibilities presented to the user, whereas the value of ir_{min} is the corresponding minimum interest density of the test instance in any subspace presented to the user. The variable PATH consists of the pointers to the sequence of successively reduced training data sets which are obtained in the process of interactive decision tree construction. We initialize the list PATH to a single element which is the pointer to the original data set $\mathcal{D}$. At each point in the decision path construction process, we determine the subspaces $E_1 \ldots E_q$, which have the greatest interest density (of the dominant class) in the locality of the test instance t. This process is accomplished by the procedure *ComputeClassifSubspaces* and is described in detail in a later section. Once these subspaces have been determined, the density profile is constructed for each of them by the procedure *ConstructDensityProfile*. We note that even though one subspace may have higher interest density at the test instance than another, the true value of a subspace in separating the data locality around the test instance is often a subjective judgement which depends both upon the interest density of the test instance and the spatial separation of the classes. Such a judgement requires human intuition which can be harnessed with the use of the visual profile of the accuracy density profiles of the various possibilities. These profile provides the user with an intuitive idea of the class behavior of the data set in various projections. If the class behavior across different projections is not very consistent (different projections are indicative of different classes), then such a node is not very revealing of valuable information. In such a case, the user may choose to back track by specifying an earlier node on PATH from which to start further exploration.

On the other hand, if the different projections provide a consistent idea of the class behavior, then the user utilizes the density profile in order to isolate a small region of the data in which the accuracy density of the data in the locality of the test instance is significantly higher for a particular class. This is achieved by the procedure *IsolateData*. This isolated region may be a cluster of arbitrary shape depending upon the region covered by the dominating class. However, the use of the visual profile helps to maintain the interpretability of the isolation process in spite of the arbitrary contour of separation. A detailed description of this process will be provided in a later section. The procedure returns the isolated data set $\mathcal{L}'$ along with a number of called the *accuracy significance* $p(\mathcal{L}', \mathcal{C}_i)$ of the class $\mathcal{C}_i$. The pointer to this new data set $\mathcal{L}'$ is added to the end of PATH. At that point, the user decides whether further exploration into that isolated data set is necessary. If so, the same process of subspace analysis is repeated on this node. In the following subsections, we will provide further descriptions of the individual procedures in decision path construction.

3.1 Determination of Subspace Alternatives

In the previous section, we discussed how the kernel density method can be used in order to analyze the discrimination behavior of the data in different subspaces. In this section, we will discuss the actual details of the procedure *ComputeClassifSubspaces*. This determines the alternative subspaces at a given node. The input to the procedure is the test instance t, the data set $\mathcal{L}$, the branch factor b_{max}, the minimum interest density ir_{min}, and the maximum dimensionality l of the subspaces in which the classification behavior is to be determined. Since, we utilize visual profiles in order to provide the user an understanding of the data, the natural choice of l is 2, though higher dimensional subspaces containing significant classification patterns are also discovered by the sequence of hierarchical decisions made by the user.

The overall subspace determination procedure uses a roll-up mechanism analogous to [1] in which the accuracy density is computed at the test instance t in different subspaces. The set of discriminatory subspaces is maintained in $\mathcal{F}$. The set of all k-dimensional candidate projections is denoted by $\mathcal{S}_k$. The value of $\mathcal{S}_1$ is the set of all 1-dimensional projections $\{1, \ldots d\}$. The algorithm starts by initializing $\mathcal{F}$ to the b_{max} 1-dimensional subspaces which have the highest interest density at the test instance t. In each iteration, the set $\mathcal{S}_k$ is generated from the set $\mathcal{S}_{k-1}$ in an iterative mechanism in which we join the candidates in $\mathcal{S}_i$ with the set of all singleton subspaces in $\mathcal{S}_1$. Those subspaces SS which have interest density $\mathcal{IM}^{SS}(t, \mathcal{L})$ higher than any subspace in $\mathcal{F}$ are retained. This process is continued until we determine the (at most) b_{max} most discriminatory subspaces with dimensionality at most l, and interest density greater than ir_{min}. For lower dimensionalities such as $l = 2$, the above procedure can be executed relatively efficiently since the density needs to be calculated only at the test instance t. This can be achieved in a single scan of the data in order to calculate the additive kernel density value at t for the different 1-dimensional and 2-dimensional candidate subspaces. We further note that even though each node contains only 2-dimensional candidates, the final combination of dimensions is determined by the path decided by the user.

3.2 Construction of Visual Density Profiles

Once the most discriminatory subspaces have been determined at a given node, we construct the visual profile of the accuracy density in these projections in the procedure *ConstructDensityProfile*. We proceed in two steps. In order to construct the visual profile, we compute the accuracy of the dominant class at a set of discrete grid points. Each attribute range is divided into a set of ϕ intervals. The intersection of these interval divisions form the grid points at which the accuracy density values are computed. The surface plot of these density values provides a good idea of the regions most closely related to the test instance which have high accuracy density. An example of such a profile is illustrated[1] in Figure 2(a). It is clear that the test instance is located in a region with high elevation of the accuracy profile. This is because the subspace was specifically selected in order to expose those subspaces in which the test instance has high accuracy.

However, the real value of a subspace can only be judged based on the behavior of the entire profile and the spatial separation of this elevation with other regions. This is because the accuracy or interest density at t does not provide a complete understanding of the class behavior with respect to the remaining data set. For example, consider the subspaces E_1 and E_2. The subspace E_1 may have a higher value of $\mathcal{IM}^{E_1}(t, \mathcal{L})$ ($\mathcal{AM}^{E_1}(t, \mathcal{L})$) than the corresponding value $\mathcal{IM}^{E_2}(t, \mathcal{L})$ ($\mathcal{AM}^{E_2}(t, \mathcal{L})$). However, the true discriminatory power can only be distinguished by using the overall spatial distributions of the different classes. In such cases, the intuition of the human is very useful in determining whether a small data locality around the test instance shows a significantly higher accuracy density than the remaining data. In such cases, it may be desirable to isolate the particular data locality for further exploration.

[1]In all future accuracy profiles, we will assume that the class for which the accuracy density is displayed is the dominant one.

3.3 Isolation of Local Data Segments

An easy way of isolating smaller segments of the data is for the user to specify an accuracy density threshold λ. Such a choice can be made relatively easily by the user, once the visual profiles of the data are directly available to him. All data points in the locality of the test instance t which have accuracy density above λ can then be utilized in order to determine the visual profile. We note that even though the choice of the data isolation parameter λ depends upon the user, it should be chosen such that the interest ratio of the accuracy value of λ should be at least 1. Therefore, if the data set $\mathcal{L}$ contains a total of $|\mathcal{L}|$ points out of which α belong to the dominant class, then the default value of λ is given by $\alpha/|\mathcal{L}|$. The actual value of λ may then be modified by the user so that a well defined local region around the test instance can be clearly distinguished. Such a judgement can be effectively made only by human perception and intuition. This is one of the advantages of an exploratory approach which is able to incorporate human feedback and intuition into the classification process.

In order to find the data points in the locality of the test instance which have accuracy density above λ, we define the concept of *accuracy connectivity* between two data points. Let $Q(t, E)$ be the dominant class at the test instance t in subspace E. We assume that the accuracy density of this dominant class at test instance t in subspace E is above the user specified threshold λ.

DEFINITION 3.1. *A data point $x \in \mathcal{D}$ is said to be accuracy connected to test instance t in subspace E at threshold λ, if there exists a path P connecting x to t, such that for any point $y \in P$, $\mathcal{A}^E(y, Q(t, E), \mathcal{D}) \geq \lambda$.*

In Figure 2(b), we have imposed an accuracy threshold by utilizing a hyperplane which is superposed on the accuracy density profile. In this case, the tiny island containing the test instance is the region from which the relevant data set is isolated. We note that even though the visual representation is quite interpretable, the minimum bounding rectangles of the various isolated regions may be used in order to provide a rough idea of the relevant parameters in the corresponding dimensions. This does not compromise the quality of the exploration process, since the actual isolation of the data is done using the arbitrarily shaped regions, whereas the use of MBRs is only a posterior step in order to maximize interpretability.

We use the grid discretization of the data in order to make an approximate computation of whether the test instance is accuracy connected to the data points. This discretization automatically separates out the subspace into rectangular regions. A grid rectangle is said to have accuracy larger than λ if the center of the rectangle has accuracy density larger than λ. The first step is to find the unique grid rectangle containing the test instance. Starting from this rectangle, we keep searching for adjacent rectangles with accuracy density above the threshold λ, until all such rectangles have been determined. Two rectangles are said to be adjacent if they share one side. Finally, all the data points that lie in these rectangles are returned as the data points which are locally relevant to the test instance t, and are used for further exploration in subsequent iterations of the *SubspaceDecisionPath* algorithm.

The procedure also returns the *accuracy significance* $p(\mathcal{L}', \mathcal{C}_i)$ for the class $\mathcal{C}_i$ of the user-defined split. Let l_i and l_i' be the

number of instances of class $\mathcal{C}_i$ in $\mathcal{L}$ and $\mathcal{L}'$ respectively. Then, we define the accuracy significance of the isolated data set $\mathcal{L}'$ with respect to class $\mathcal{C}_i$ as follows:

$$s(\mathcal{L}', \mathcal{C}_i) = (l_i'/|\mathcal{L}'| - l_i/|\mathcal{L}|)/\sqrt{(l_i/|\mathcal{L}|) \cdot (1 - l_i/|\mathcal{L}|)/|\mathcal{L}'|} \quad (6)$$

We note that this significance factor is the number of standard deviations by which the class fraction of $\mathcal{C}_i$ in the isolated data set $\mathcal{L}'$ is larger than the data set $\mathcal{L}$. Correspondingly, the significance factor of class $\mathcal{C}_i$ is defined as follows:

$$p(\mathcal{L}', \mathcal{C}_i) = \max\{0, 2 \cdot \Phi(s(\mathcal{L}', \mathcal{C}_i)) - 1\} \quad (7)$$

Here $\Phi(\cdot)$ is the cumulative normal distribution function. The value of p is non-zero only when the isolated data set $\mathcal{L}'$ contains a larger proportion of class $\mathcal{C}_i$ than the original data set $\mathcal{L}$. By approximating the significance factor with a normal distribution, we are able to quantify the probabilistic level of significance that the isolated data set is significantly more indicative of a particular class.

3.4 Termination

Since the aim of this paper is to provide the user with open-ended exploratory ability, the final decision of termination is dependent upon the user. At the same time, we wish to provide the user with some intuitive guidance as to when termination should take place. In this section, we will discuss the behavior of the visual profiles as well as some statistical measures which provide evidence of termination. We note that isolation of data localities results in successive refinement such that the locality around the test instance is increasingly dominated by a particular class. This can also be perceived in the visual density profile of the data at the lower levels of the decision path. In such cases, the test instance occurs on a plateau of high accuracy density in the visual profile. Examples of such accuracy density profiles are illustrated in Figures 2(c) and 2(d). In these cases, the test instances have accuracy densities 98.0% and 99.0% respectively in the corresponding projections. It is also clear from Figures 2(c) and 2(d) that the region in the immediate locality of the test instance is flattened out at the accuracy density of the test instance. This behavior is the result of successive data isolations which are akin to a visual magnification of the data locality with the use of carefully chosen subspaces. Our empirical results illustrate the interesting phenomenon that even though only the small number of subspaces on the decision path are used for isolation of data locality, all the subspaces on the lower nodes of the PATH start exhibiting this behavior. This is because of the inter-attribute correlations of the different attributes all of which provide consistent information about the class behavior.

A better statistical measure of the level of significance of a given path is obtained by computing the cumulative dominance level of each class $\mathcal{C}_i$ along PATH. Let $\mathcal{L}_0 \ldots \mathcal{L}_r$ be the nodes along PATH. Then, we define the cumulative dominance $CD(PATH, \mathcal{C}_i)$ for the class i as follows:

$$CD(PATH, \mathcal{C}_i) = 1 - \pi_{j=1}^{r}(1 - p(\mathcal{L}_j, \mathcal{C}_i)) \quad (8)$$

The larger this value, the greater the cumulative dominance of $\mathcal{C}_i$ along PATH. A user may choose to stop exploration along this path, when the cumulative dominance of some class $\mathcal{C}_i$ exceeds a pre-defined threshold. Unlike a decision tree, the user may also traverse multiple paths of the decision process by back-tracking instead of terminating when a path has been successfully explored.

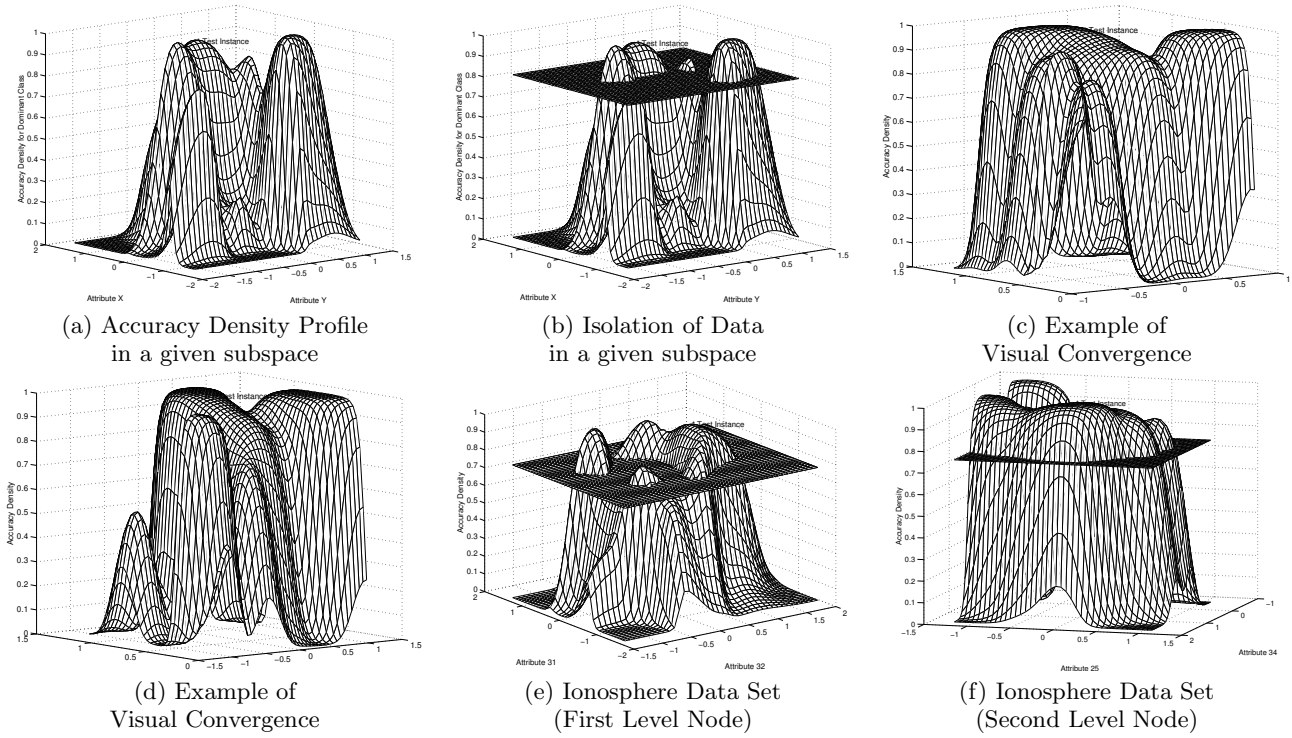
Table 1: Accuracy of SD-Path Method

Data Set	Correct (SDP)	Indeter. (SDP)	Incorr. (SDP)	C4.5 (Corr.)
Ionos.	20	0	0	18
Segmen.	19	1	0	16
Glass	19	1	0	17
Ecoli	19	1	0	16

4. EMPIRICAL RESULTS

In this section, we will provide a detailed discussion and understanding of the advantages of the instance-centered exploratory approach developed in this paper. The ionosphere data set from the UCI machine learning repository contains 34 attributes and two classes corresponding to "g" or "b" depending upon the quality of radar returns from the atmosphere. In Figure 2(e), we have illustrated the accuracy density of the dominant class label in the highest interest density subspace. The corresponding accuracy densities of the test instance ranged from between 93% to 96% for all the different subspaces found. All of the branches had the same dominant class label corresponding to "g". Figure 2(e) illustrates the visual profile of one of these branches, which shows particularly high level of discrimination in its data locality. The accuracy density value of the dominant class label at the test instance is 95.8%. In order to obtain a further idea of the local behavior of the data around this test instance, we decided to explore into the data locality which was accuracy connected to the test instance at a threshold of 75.0%. The isolated data was the island region containing the test instance in Figure 2(e). Note that this region has a somewhat irregular shape which cannot be characterized easily in closed form, but could be understood more easily in this visual representation. Upon expanding this region further, we found that all the branches again corresponded to "g" with accuracy density values between 98% and 99%. An example of such a branch is illustrated in Figure 2(f). An interesting characteristic of this profile is that the region in the immediate locality of the test is somewhat flattened out at a high accuracy rate. This kind of behavior is true for all branches at that level. This shows that the process of successive data isolations has resulted in a smaller and more refined locality around the test instance in which the user can clearly perceive a consistently high concentration of a particular class. If all of the projections corresponding to the different branches exhibit this behavior for the same class $\mathcal{C}_i$, then it is likely that the test instance belongs to $\mathcal{C}_i$.

In order to determine the effectiveness of the SD-Path method, we also tested the overall accuracy of the approach. In each case, we predicted a class label, if it was found to converge to the same value using three separate decision paths. If the class label was found to be inconsistent on three separate paths, then we labeled the corresponding test instance as indeterminate. we further note that while we also present the results in comparison to the traditional C4.5 classifier, it is important to understand that our classifier is *not* an alternative to traditional batch classifiers which can classify a large number of test instances. Rather, it is only intended as a *diagnosis* tool for individual test instances. This is particularly useful for test instances containing conflicting classification charactersitics in different dimensions. Therefore, the only purpose of these comparisons is to illustrate that



(a) Accuracy Density Profile in a given subspace

(b) Isolation of Data in a given subspace

(c) Example of Visual Convergence

(d) Example of Visual Convergence

(e) Ionosphere Data Set (First Level Node)

(f) Ionosphere Data Set (Second Level Node)

Figure 2: Results of SD-Path Method

the exploratory classifier is consistently more robust than a traditional classifier because of the use of human intervention in the process.

In Table 1, we have illustrated the accuracy of the SD-Path method on a number of data sets from the UCI machine learning repository. In each case, we ran the exploratory SD-Path approach over a set of 20 examples and found that in most cases, the SD-Path method was able to obtain consistent results over the different paths. In the same table, we have also illustrated the effectiveness of the C4.5 method. It is clear that the SD-Path method is much more accurate than the C4.5 technique. This is partially because of the fact that the SD-Path method is able to exploit both the local behavior of the test instance as well as the intuition of the user during the classification process. Another interesting observation from Table 1 is that in each case, none of the test instances were classified inaccurately, though some were considered indeterminate by the SD-Path method. The reason for this indeterminate behavior was the fact that the test instances shared characteristics from multiple classes. This is evidence of the anamolous behavior of the test instance rather than a limitation of the SD-Path method. In fact, this diagnosis helps the user understand the various combinations of dimensions which reveal this contradicting behavior.

5. CONCLUSIONS AND SUMMARY

In this paper, we discussed the subspace decision path method, an effective exploratory instance-based approach for decision path construction for high dimensional data sets. The advantage of this method is that it effectively combines the data mining process with human interaction in order to provide good understanding of the classification characteristics of a given test instance. Since the process uses test-instance specific local subspace characteristics of the data, it is much more flexible and concise than a decision tree construction process. Our empirical tests illustrate that this flexibility also results in a significantly more accurate classification process. The ability to explore multiple paths of an instance-specific process provides the user with multiple perspectives of the important characteristics in the instance. Even in cases where the classification behavior of the instance is poorly defined, the subspace decision path method is able to provide insight into the different characteristics of the test instance which have contrasting behavior. Such information is of great value in a number of business applications in which the causality of the classification process provides valuable information to the end-user.

6. REFERENCES

[1] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. *VLDB Conference*, 1994.

[2] M. Ankerst, M. Ester, H.-P. Kriegel. Towards an Effective Cooperation of the Computer and the User for Classification. *KDD Conference*, 2000.

[3] B. Liu, W. Hsu, Y. Ma. Integrating Classification and Association Rule Mining. *KDD Conference*, 1998.

[4] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.