Defining *diff* as a data mining primitive

Ramesh Subramonian,
MicroComputer Research Laboratory,
Intel Corporation,
2200 Mission College Blvd, MS RN6-35
Santa Clara CA 95052-8119
subramon@gomez.sc.intel.com

Abstract

The emphasis on discovery in the knowledge discovery process while important in its own right, has distracted from the equally important process of knowledge representation and maintenance. For a system to indicate what is new or different, it must have an understanding of what is old or well understood or expected.

In this paper, we propose *diff* as a fundamental data mining primitive. We show how it can be used to capture knowledge, either as a set of representative instances or as a set of rules, in a framework that is tightly integrated with the knowledge discovery process. We show how it can be applied to both discrete and continuous attributes and association rules. Lastly, we show how it enables the user to pinpoint high-level differences between two data sets that share the same attributes.

Introduction

The knowledge discovery process has been defined as "the nontrivial process of identifying valid, **novel**, potentially useful, and ultimately understandable patterns in data" (FPSS96). However, what is sometimes ignored is that the ability to determine what is *new* is predicated on the ability to differentiate it from what is *old*. We believe the ability to model, maintain and incrementally update an existing knowledge base in a manner that is tightly integrated with the discovery process is key to the successful deployment of data mining techniques on a wide spread basis. In this paper, we propose the diff operator which addresses the problem that to tell you what is new, the system needs to understand what you perceive to be old.

The observation that many of the association or implication rules produced by market basket analysis are obvious (BMUT97) has led to attempts to reduce the number of rules produced by metarule-guided mining (KHC97) and the specification of filters (KMRT94) and finding exceptions (Suz97).

To use a marketing example, it is well known that most PCs made in the US are made by worldwide suppliers, whereas most PCs made in Brazil are made by mom-and-pop shops. While one can conceive of a data mining system discovering this fact, what is more important is that it maintain this knowledge and alert the user if the relative percentage of manufacturers in a given geographic region changes by more than a certain amount. Beliefs about the behaviour of attributes can be collected as individual profiles and/or as group profiles, allowing for collaborative learning.

Data is often collected on an incremental basis (weekly, monthly, ...). When new data is collected, a key question to ask is how it differs from existing data e.g., how do sales in 1996 differ from sales in 1997 ? We provide a mechanism to highlight the key differences between data sets that share the same meta data.

The aim of this paper is to define a problem of interest and to provide a framework for its solution. The contributions of this paper are :
1. define diff as a data mining primitive
2. show how diff unifies the treatment of existing data mining techniqes
3. show how a knowledge base can be constructed to encode the user's beliefs and focus subsequent reports on new findings
4. provide experimental validation of the usefulness of diff.

Background

The purpose of diff is:

1. to detect behaviour that deviates from the norm, where the norm can be defined in terms of a set of examples or a set of rules.
2. when presented with two data sets that have been measured on the same set of attributes, to rank the attribute combinations in their ability to highlight the differences between the data sets.

We define diff in terms of δ and ϵ, which are similar to the "confidence" and "support" parameters used to define association rules (AIS93). The δ parameter measures the extent of the difference and the ϵ parameter

measures the frequency of occurrence of this difference.

For simplicity and uniformity of notation, we shall use probability density functions (pdf) for problem definitions. In practice, we anticipate that pdfs will have to be approximated by counting the frequency of occurrence of the event whose probability is being estimated (as is done in association rules (AIS93)) or density estimation techniques (Sil86). Given n variables $X_1, \ldots X_n$, we assume that the joint pdf and all relevant marginzaliations thereof can be estimated from the data i.e., $P(X_i), P(X_i, X_j), \ldots P(X_1, \ldots X_n)$. To the extent that data insufficiency or computational considerations make calculations of these pdfs infeasible, the algorithms proceed as far as time and data permit. The pdf of a continuous variable is sampled finely so as to unify its treatment with the discrete case.

Since the entire purpose of diff is to look for things that are different in a probabilistic sense, we need to define what we mean by dissimilar. Two alternative definitions of the difference between and inequality of probabilities, p_1, p_2, are provided in Definitions 0.1, 0.2.

Definition 0.1 *Let* $D(p_1, p_2) = |p_1 - p_2|$. $p_1 \neq_\delta p_2$ *if* $D(p_1, p_2) > \delta$; *else,* $p_1 =_\delta p_2$.

Definition 0.2 *Let* $D(p_1, p_2) = max(\frac{p_1}{p_2}, \frac{p_2}{p_1})$. $p_1 \neq_\delta p_2$ *if* $D(p_1, p_2) > \delta$; *else,* $p_1 =_\delta p_2$.

The distance between two pdfs p and q that share the same range can be measured using a variety of metrics such as the Kulback-Leibler distance $\sum_x p(x) \log \frac{p(x)}{q(x)}$, the L-1 norm $= \sum_x |p(x) - q(x)|$, etc. Let D be such a metric. Definition 0.3 defines inequality of pdfs.

Definition 0.3 $p(x) \neq_\delta q(x)$ *if* $D(p(x) \| q(x)) > \delta$

Discrete variables

We shall use the simple case of unordered, discrete variables to lay the groundwork for the definition of diff. For brevity, we shall use the least number of variables to define the problem statement with the implicit understanding that problems defined in terms of $P(X)$ can easily be extended to those involving $P(X, Y)$ and so on.

Difference between data sets Given a set of representative instances, we often wish to determine how and where a new set of instances differs from it. Let X be an attribute in the data set, which we model as a random variable. Let $p(X)$ and $q(X)$ be the estimated pdf of X on the two different data sets. The diff operator returns all values x' of X where $p(X = x') \neq_\delta q(X = x')$ and $max(p(X = x'), q(X = x')) > \epsilon$.

Problem Definition 0.1 *Given two joint pdfs, p and q, of n variables* $X_1, \ldots X_n$, *and all pertinent marginalizations thereof, find all k-tuples of the form* $\bar{X} = [X_{i_1} = x'_{i_1}, \ldots, X_{i_k} = x'_{i_k}]$ *such that* $p(\bar{X}) \neq_\delta q(\bar{X})$ *and* $max(p(\bar{X}), q(\bar{X}) > \epsilon$.

Difference as a correlation indicator The possible existence of a correlation can be found by an alternative formulation of the diff operator. There are several levels at which this can be determined. At the highest level, we can use the correlation or the mutual information between X and Y. Since these are gross metrics which might conceal subtle variations, we need to examine them at finer levels. We do this by using the fact that two random variables X and Y are independent if $P[X|Y] = P[X]$ and hence looking for cases where this equality is violated in a substantial manner.

Problem Definition 0.2 *Given* $p(X)$ *and* $p(X|Y)$, *find all pairs of values of X and Y,* (x, y), *where* $p[X = x] \neq_\delta p[X = x|Y = y]$ *and* $p[X = x] > \epsilon$.

An alternative mechanism to focus on causality is as follows. Within a particular attribute, Y, we are interested in learning about pairs of values, y_1, y_2 which differ greatly in their influence on a related attribute X. Running this on the adult data set (MM96) provides rules of the form $p_1 = P[> 50K|\text{Preschool}] \neq_\delta P[> 50K|\text{Doctorate}] = p_2$ [1] Since such a rule might well be considered obvious, what would be more interesting is to ask the system to inform us when this is not the case, as is done in Section .

Problem Definition 0.3 *Given* $p(X)$ *and* $p(X|Y)$, *find all triples* (x_i, y_j, y_k), *where* $p[X = x_i|Y = y_j] \neq_\delta p[X = x_i|Y = y_k]$ *and* $max(p[X = x_i|Y = y_j], p[X = x_i|Y = y_k]) > \epsilon$.

Noise reduction. Data mining is plagued with the creation of too many, often irrelevant, redundant or obvious, rules (BMUT97). Agglomeration of values (with user controlled drill down) is one mechanism to reduce the number of rules. This can be done manually by the creation of derived attributes e.g., the grouping of manufacturers such as $\{Compaq, Dell, NEC, \ldots\}$ into a single value, *Tier 1.* It can also be done automatically by finding maximal sets $\{y_j\}$ and $\{y_k\}$ for which $P[X = x_i|Y \in \{y_j\}] \neq P[X = x_i|Y \in \{y_k\}]$, rather than reporting all $|\{y_j\}| \times |\{y_k\}|$ rules of the form $P[X = x_i|Y = y_j] \neq P[X = x_i|Y = y_k]$.

Knowledge Base construction

A knowledge base (kb) is needed to capture what the user is interested in, what the user is not interested in, what the user knows, what the user believes to be true. While it may be argued that the construction of a kb is tedious, we believe that it is not onerous because (i) it can be constructed incrementally (ii) it is domain

[1] In this case, $p_1 = 0, p_2 = 0.74, D(p_1, p_2) = 0.74$ by Definition 0.1, $D(p_1, p_2) = \infty$ by Definition 0.2

specific and real users tend not to flit from domain to domain (iii) it is possible to construct short-hand notations which can specify interest/disinterest/beliefs in a large number of events in a concise manner. While different entries in the *kb* take on different forms depending on their structure, at a high level, an entry in the *kb* is a belief specified in terms of an event and a range of acceptable probabilities for that event. Equivalently, since not all valid statistical phenomena are of interest, the knowledge base also consists of phenomena that are *not* of interest. Events can be denoted explicitly or implicitly. The following represent ways in which the *kb* can be populated.

1. Explicit events. These are typically the outcome of previous data mining sessions. Every output of diff is a potential candidate for inclusion in the *kb*. To use our marketing example, an entry of the form (*[Manufacturer=Tier 1|Country =US]*, 0.9, 0.1) would alert the user if the market share of Tier 1 manufacturers in the US fell below $80\% = 0.9 - 0.1$. More complex events can be specified visually e.g., Figure 2, where the complement of the region with the thick border represents unlikely mpg, displacement combinations.

   Alternatively, *kb* entries could be preconceived notions of the form (*correlation*(X, Y), 0, δ) indicating our belief that X and Y are independent. As an example, we were surprised to find a correlation, albeit small, between *cylinders* and *model-year* in the auto data set (MM96). A plot of the joint density (Figure 1 led us to the explanation: the oil crisis of the early seventies encouraged a gradual trend towards high mileage cars.

2. Implicit events. These encompass association rules (AIS93), implication rules (BMUT97), negative associations and exception rules (Suz97) (Section).

3. Don't care entries. These represent a set of events that the user is uninterested in. As an implementation convenience, these can be specified at various levels of detail. Some of the short-hand notations we propose are:
   (a) (X, Y): all events of the form $[X = x_i, Y = y_i]$
   (b) $(X|Y)$: represents all events of the form $[X = x_i|Y = y_i]$
   (c) $(X, *|Y, *)$: represents all events of the form $X = x_i, \ldots |Y = y_i, \ldots]$
   (d) (correlation, X, Y): represents all pair-wise correlations

4. Representative instances and a trigger δ. For the entire set of diff operations e.g., (i) $D(p(X)\|q(X))$, (ii) $[X = x, Y = y, \ldots]$, (iii) $X = x|Y = y]$, the representative set is one way of specifying the associated probabilities for a large number of events. Given a new data set, all significant (measured by ϵ) events that differ (measured by δ) from the corresponding events calculated on the representative set are flagged.

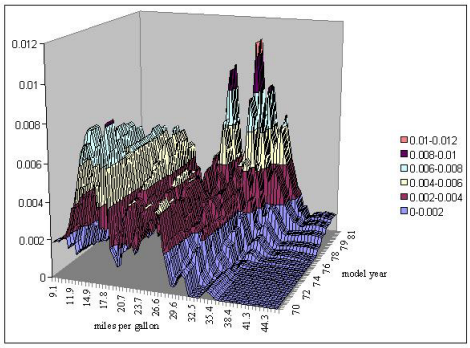

Figure 1: Plot of mpg and model-year in auto data set

Diff and association rules

In this section, we show how association rules (AIS93) and implication rules (BMUT97) can be the cast in the diff framework as short-hand instantiations of the kb with implicit rules.

Let there be N items $X_1, \ldots X_n$ which can be treated as random variables with outcomes $\{Y, N\}$ depending on whether or not they occur in a transaction. A transaction is a random variable with 2^n outcomes in $X \in \{X_1 = Y, X_1 = N\} \times \ldots \{X_n = Y, X_n = N\}$. The association rules problem is typically stated as wanting to find all rules of the form $P[X_1|X_2] >$ *confidence* $\wedge\, P[X_2] >$ *support*. This is equivalent to populating the kb with the implicit beliefs that comprise the (event, probability, trigger) triple of the form $[X_{1,1} = Y, \ldots X_{1,n_1} = Y|X_{2,1} = Y, \ldots X_{2,n_2} = Y]$, 0, *confidence* and $x \neq_\delta y$ is defined as in Definition 0.1. The user is notified of events whose estimated probability is different from 0 by more than *confidence*. The role of *support* is to ensure robust probability estimation.

Casting the notion of *interest* defined in (BMUT97) in the diff framework assumes an implicit knowledge base containing events of the form $[X_{i_1}, X_{i_2} \ldots X_{i_{n_1}}]$ with an associated probability of $P[X_{i_1}] \times P[X_{i_2}] \ldots P[X_{i_{n_1}}]$. This expresses our belief that we assume the events $X_{i_j} = Y$ to be independent and wish to be notified if the contrary is discovered.

Similarly, their notion of "conviction" assumes an implicit knowledge base of the form $[X_{1,1} = Y, \ldots X_{1,n_1} = Y|X_{2,1} = Y, \ldots X_{2,n_2} = Y]$ and an associated probability of $P[X_{1,1} = Y] \ldots P[X_{1,n_1} = Y]$. In other words, the occurrence of $[X_{2,1} = Y, \ldots X_{2,n_2} = Y]$ significantly alters the probability of the event $[X_{1,1} = Y, \ldots X_{1,n_1} = Y]$.

To use an example from (BMUT97), they find that the conviction of *Vietnam veteran* $\Rightarrow$ *more than five years old* is ∞. In our framework, using Definition 0.2, $D(P[$ *Vietnam vet* $|$ *less than five years old*$], P[$ *Vietnam vet*$]) = \infty$.

Continuous variables

The treatment of continuous variables differs from discrete variables (Section) in that the infinite [2] values that the variable can take on and the ordering among them both requires and encourages us to group values in order to find regions that satisfy the ϵ requirement. Consider a single attribute X for which pdfs p and q have been estimated from two data sets. The diff operator returns all ranges $[x_j, x_k]$ of X where $\forall i,\ j \leq i \leq kp(x_i) \neq_\delta q(x_i)$ and $max(\sum_{i=j}^{i=k} p(x_i), \sum_{i=j}^{i=k} q(x_i)) > \epsilon$. In order to reduce the number of such ranges found, we require the ranges to be maximally different (Definition 0.4).

Definition 0.4 *Let X be a continuous-valued random variable and $p(X)$ be an estimate of the pdf of X, evaluated at n locations, $\{x_1, x_2, \ldots x_{n_X}\}$ where $x_i < x_{i+1}$. A region, $[x_i, x_j]$, is said to be maximally different if $\forall i,\ j \leq i \leq k\ p(x_i) \neq_\delta q(x_i)$ and $p(x_{j-1}) =_\delta q(x_{j-1})$ and $p(x_{k+1}) =_\delta q(x_{k+1})$.*

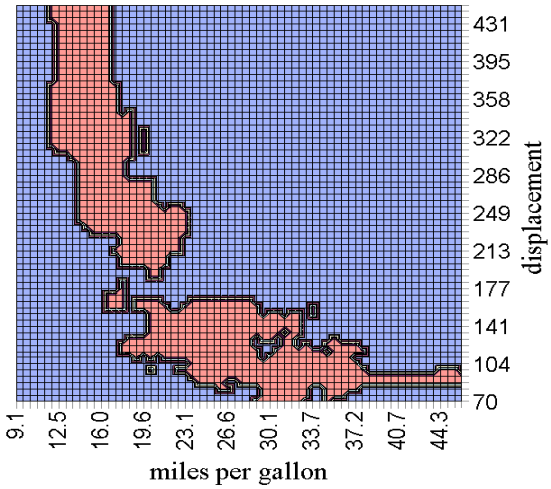

Figure 2: Thresholded pdf of mpg and displacement in auto data set

Multi-dimensional analysis is motivated by the fact that patterns that do not exist in fewer dimensions are thrown into sharp relief in higher dimensions. However, it is complicated by the inability to impose a linear ordering on the event space and hence a simple equivalent of Definition 0.4. In this case, definition of maximal subsets of events can be done by user intervention (supported with appropriate visualizations) or likelihood-based definitions as in Figure 2. The figure shows the smallest region that encompasses some specified fraction of the instances, 97% in this case. Visually defining such a region enables

1. creation of a *probabilistic query*. Databases need to be able to support such "regions of interest" queries efficiently.

2. *probabilistic integrity constraints.* Databases need to extend their notion of integrity constraints to be able to identify not just that which is impossible (e.g., age < 0) but what is improbable. For example, the placement of the complement of the 97% region of Figure 2 in the knowledge base allows the identification of the unexpected. In this case, an Oldsmobile Cutlass with a 262 cu. in. engine and a mileage of 38 stands out as a diesel car. Note that there is no single attribute value in which this diesel car stands out from the rest.

Implementation Detail. Computational conveniences are possible, by restricting the application of diff to the paremeters, when continuous random variables are modelled using parametric forms e.g., Gaussians.

Difference between clusters

Clustering algorithms (CS96) can benefit from the application of diff in the following ways.

1. determine whether the cluster that best describes a new data set is substantially different from that for an old data set.

2. clustering is a computationally intensive search process, often implemented as several sequential searches with different initial conditions. Hence, it is amenable to an incremental/anytime (SW97b) approach, where the results of one search are presented while other searches are outstanding. Diff is used to alert the user to the discovery of a significantly better/different cluster.

The difference between clusters is measured, at various levels, in the following ways

1. Let $p_1(J)$ be the pdf representing our belief that data set 1 is best described by $J = j$ classes. $D(p_1 \| p_2)$ measures whether data sets 1 and 2 are best described by the same number of classes.

2. Let $L_1 = \sum_{i=1}^{i=I} \log(P[X_i|C_1])$ be the log-likelihood of a set of I instances $\{X_i\}$, given a cluster C_1. Clusters C_1 and C_2 differ in their ability to describe the data if $L_1 \neq_\delta L_2$.

3. Let $\bar{X}$ be a point in the K-dimensional space being clustered. Define $p_1(\bar{X}, j)$ as the probability that $\bar{X}$ is assigned to class j in cluster C_1. Two clusters, C_1 and C_2, are considered different ($C_1 \neq_\delta C_2$) if $D(p_1 \| p_2) > \delta$. However, a complication that we have glossed over is that class j_1 of cluster 1 may actually be most like class j_2 of cluster 2. Let $\pi : \{1, \ldots J\} \rightarrow \{1, \ldots J\}$ be a permutation. Let $p_2^\pi(\bar{X} = j) = p_2(\bar{X} = \pi(j))$. Taking class number permutations into account, $C_1 =_\delta C_2$ if there exists some permutation π such that $D(p_1 \| p_2^\pi) < \delta$. From a practical standpoint, it is often easier to measure the differences in the parameters of the functional forms rather than pdfs.

[2] Recall that to enable numerical integration, X has been finely discretized into n_X values for some large n_X.

Ranking Differences

When applied to two data sets that share the same meta data, "diff" returns a high-level description of how and where the data sets differ. We do this by ranking combinations of attributes in terms of the distance between the pdfs estimated for those attributes from the two different data sets. The problem can be stated as: given pdfs $p(X_i)$ and $q(X_i)$ for n variables X_i, find the k largest differences $D(p(X_i)\|q(X_i))$ and the corresponding attributes. Similarly for joint pdfs of 2 or more variables. Comparing Males versus Females in the adult data set (MM96) showed that the pair-wise attribute combination in which the data sets differed the most was *occuptation* and *marital-status*. Drilling down into this pair (Figure 3) shows a large fraction of female unmarried clerical workers and a large fraction of male married craftsmen. Various metrics could be proposed to measure the differences between the pdfs. Our experiments with the KL-Distance and L-1 norm (details omitted) indicate that the ranking is not very sensitive to the choice of metric

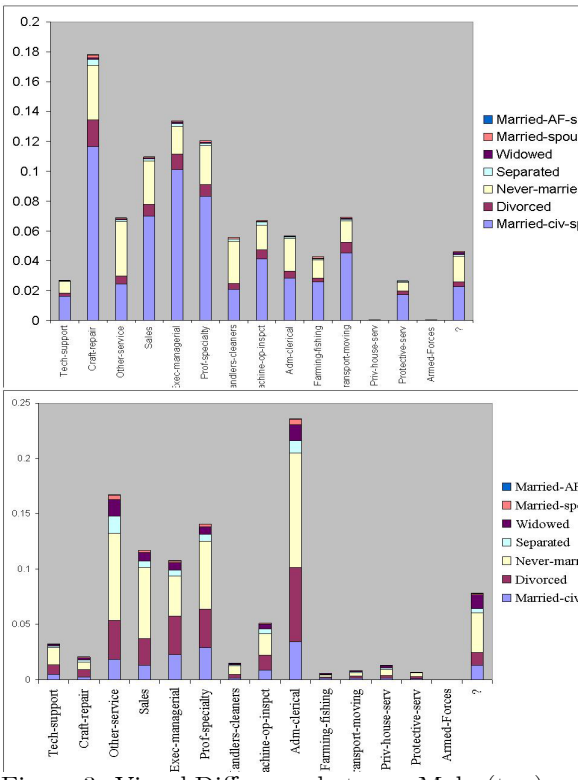

Figure 3: Visual Difference between Males(top) and Females (bottom) adult data set

Conclusion and Future Work

We have proposed diff as a fundamental data mining primitive that captures the user's beliefs in terms of events, associated probabilities and triggers. These are defined by a set of representative instances and/or implicitly or explicitly user-specified rules. We have shown how it can be deployed in a variety of settings, including specification of probabilistic integrity constraints on continuous and discrete attributes, association rules and clustering. We have shown how one can visually compare two data sets to find how they most differ, as specified by a combination of attributes along which the maximum difference is observed. Diff represents our belief that the knowledge discovery process can be, and ought to be, integrated tightly with the knowledge maintenance and representation process. The design of benchmarks and algorithms should reflect the reality that the KDD process, like data acquisition, is an ongoing, not a one-time operation.

Diff throws open a wide variety of issues that need to be addressed for its efficient implementation. We have identified two key ways in which the database needs to support data mining. One is the specification of *probabilistic integrity constraints*. The other is *query by interest* (e.g., identifying outliers via confidence thresholds as in Figure 2).

References

R. Agrawal, T. Imilenski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp. 207–216, 1993.

S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market data. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp. 255–264, 1997.

Peter Cheeseman and John W. Stutz. Bayesian classification (Autoclass): Theory and results. In U. M. Fayyad et al, ed., *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

U. Fayyad, G. Piatesky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *CACM*, 39(11):27–34, 1996.

M. Kamber, J. Han, and J. Chiang. Metarule-guided mining of multi-dimensional association rulesusing data cubes. In *3rd Conf. Knowledge Discovery and Data Mining*, pp. 207–210, 1997.

M. Klemettinen, H. Mannila, P. Ronkainen, and P. Toivonen. Finding interesting rules from large sets of discovered association rules. In *CIKM*, pp. 401–408, 1994.

C. J. Merz and P.M. Murphy. *UCI Repository*. www.ics.uci.edu/mlearn/MLRepository.html.

Brian Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman Hall, 1986.

E. Suzuki. Autonomous discovery of reliable exception rules. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 259–262, 1997.

P. Smyth and D. Wolpert. Anytime exploratory data analysis for massive data sets. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 54–60, 1997.