CLIFF: Tools for Finding Prototypes for Nearest Neighbor Algorithms with Application to Forensic Trace Evidence

Fayola Peters

Thesis submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Computer Science

Tim Menzies, Ph.D., Chair
Arun Ross, Ph.D.
Bojan Cukic, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2010

Keywords: Forensic Evaluation, Prototype Learning, K-Nearest Neighbor

Abstract

CLIFF: Tools for Finding Prototypes for Nearest Neighbor Algorithms with Application to Forensic Trace Evidence

Fayola Peters

Prototype Learning Schemes (PLS) started appearing over 30 years ago (Hart 1968, [20]) in order to alleviate the drawbacks of nearest neighbour classifiers (NNC). These drawbacks include:

1. computation time,
2. storage requirements,
3. the effects of outliers on the classification results,
4. the negative effect of data sets with non-separable and/or overlapping classes,
5. and a low tolerance for noise.

To that end, all PLS have endeavored to create or select a *good* representation of training data which is a mere fraction of the size of the original training data. In most of the literature this fraction is approximately 10%. The aim of this work is to present solutions for these drawbacks of NNC. To accomplish this, the design, implementation and evaluation of CLIFF, a collection of new prototype learning schemes (CLIFF1, CLIFF2 and CLIFF3) are described. The basic structure of the CLIFF algorithms involves a ranking measure which ranks the values of each attribute in a training set. The values with the highest ranks are the used as a rule or criteria to select instances/prototypes which obeys the rule/criteria. Intuitively these prototypes best represents the region or neighborhood it comes from and so are expected to eliminate the drawbacks of NNC particularly 3, 4 and 5 above.

With 13 standard data sets from the UCI repository [16], the results of this work demonstrate that CLIFF presents results which are statistically the same as those from NNC. Finally in the forensic case study a data set composed of the infrared spectra of the clear coat layer of a range of cars, the performance analysis showed that is strong with near 100% of the validation set finding the right target. Also, prototype learning is applied successfully with a reduction in brittleness while maintaining statistically indistinguishable results with validation sets.

Acknowledgments

The authors would like to thank the WVU Forensics Science Initiative and Royal Canadian Mounted Police (RCMP) particularly Mark Sandercock for the glass database. This work is supported by the Office of Justice Programs, National Institute of Justice, Investigative and Forensic Sciences Division under Grant No. 2003-RC-CX-K001.

Contents

List of Figures

Chapter 1

Introduction

Since the creation of the Nearest Neighbour algorithm in 1967 (Hart [8]), a copious amount of prototype learning schemes (PLS) have appeared to remedy the five (5) major drawbacks associated with the algorithm and it's variations. First, the high computation costs caused by the need for each test sample to find the distance between it and each training sample. Second, the storage requirement is large since the entire dataset needs to be stored in memory. Third, outliers can negatively affect the accuracy of the classifer. Fourth the negative effect of data sets with non-separable and/or overlapping classes and last, the low tolerance to noise. To solve these issues, PLS are used. Their main purpose is to reduce a training set via various selection and/or creation methods to produce *good prototypes*. *Good* here meaning that the prototypes are a good representation of the original training data set such that they maintain comparable or improved performance of a nearest neighbour classifier while simultaneously eliminating the effects of the drawbacks mentioned previously.

A review of the literature on prototype learning for this thesis has yielded at least 40 PLS each indicating with experimental proof that their particular design is comparable or better than the standard schemes; published surveys of PLS can be found in [3, 4, 24]. However many of these schemes suffer from at least one of the following disadvantages:

- computationally expensive
- order effects (where the resulting prototypes are unreasonably affected by the order of the original training set)
- overfitting

The goal of this thesis is not to be unduly critical of the PLS which may succumb to any of the above disadvantages (particularly since many of them have proven to be successful), but rather to present a novel approach to this field of study which overcomes these disadvantages to report little or no loss in recognition of NNC.

Thus, this thesis presents CLIFF, a prototype learning scheme which runs in linear time, is independent of the order of the training set and avoids overfitting. A novel feature of CLIFF is that instead of either removing or adding (un)qualified prototypes from/to a 'prototype list', a conjunction of contraints is generated for each target class. These conjuctions are created using a ranking algoritm call BORE (Best Or Rest) [21], which basically finds a range of values for each attribute which best represents the specific target class i.e. have the highest ranks. Any of the instances in the training set which adheres to all the constraints is or are selected as prototypes. Using this structure the percentage reduction of the training set is related to the number of constraints used on the prototype selection process, in that the more constraints used the lower the percentage reduction. So there is no need to predetermine the number of prototypes desired when using CLIFF.

After describing the design and operation of CLIFF, its performance is demonstrated by evaluating it using cross-validation experiments with the wide variety of standard data sets from the UCI repository [16]. Our experiments also use FastMap [15] [1] and Feature Subset Selection (FSS) to avoid the curse of dimensionality [29] which negatively affects Nearest Neighbor Classifiers (NNC).

[1] A Fast Algorithm for Indexing, Data Mining, and Visualization of Traditional and Multimedia Datasets

Next we describe how CLIFF can be used as part of a tool/model for the evaluation of trace forensic evidence. The principal goal of forensic evaluation models is to check that evidence found at a crime scene is (dis)similar to evidence found on a suspect. In our studies of forensic models for evaluation particularly in the sub-field of glass forensics, we conjecture that many of these models succumb to the following flaws:

1. A tiny error(s) in the collection of data;

2. Inappropriate statistical assumptions, such as assuming that the distributions of refractive indices of glass collected at a crime scene or a suspect obeys the properties of a normal distribution;

3. and the use of measured parameters from surveys to calculate the *frequency of occurrence* of trace evidence in a population

In this work we show that CLIFF plays an effective role in the evaluation of forensic trace evidence.

Our research is guided by the following research question:

- Is CLIFF viable as a Prototype Learner for NNC?

  The goal here is to see if the performance of CLIFF is comparable or better than the plain k nearest neighbor (KNN) algorithm. So in our first experiment we compare the performance of predicting the target class using the entire training set to using only the prototypes generated by CLIFF.

1.1 Statement of Thesis

1.2 Contributions of this Thesis

1.3 Structure of this Thesis

The remaining chapters of this thesis are structured follows:

- Chapter 2 provides a survey of Prototype Learning Schemes over the years
- Chapter 3 describes the design and operation of CLIFF
- Chapter 4 describes the data sets used along with the preprocessing tools used to avoided the curse of dimensionality [29]
- Chapter 5 presents a detailed description of the experimental procedure followed to analyze the data using CLIFF
- Chapter 6 examines a case study in which CLIFF is used to reduce the *brittleness* of a forensic model
- Chapter 7 conclusions are presented

Chapter 2

Background and Related Work

2.1 Prototype Learning for Nearest Neighbor Classifiers

Research in prototype learning is an active field of study [**?**, 4, 6, 7, 9, 10, 17, 18, 25, 26, 30, 34]. A review of the literature in this field has revealed two categories of PLS: 1) instance selection, and 2) instance abstraction. Instance selection involves selecting a subset of instances from the original training set as prototypes. Using what Dasarathy terms as *edit rules*, instance selection can take place in four(4) different ways.

1. incremental (CNN [20])
2. decremental (RNN)
3. a combination of 1 and 2
4. border points, non-border points or central points

Instance abstraction involves creating prototypes by merging the instances in a training set according to pre-determined rules. For example, Chang [**?**] merges two instances if the have the same class, are closer to each other than any other instances and the result of merging does not

degrade the performance of NNC. The following section gives a brief survey of PLS for both instance selection and instance abstraction.

2.1.1 Instance Selection

Condensed Nearest Neighbor (CNN)

Reduced Nearest Neighbor (RNN)

Minimal Consistent Set (MCS)

In 1994, Dasarthy [9] presented the MCS algorithm. This is a PLS designed to select exemplar samples of a training set with the aim of reducing the computational demands of NNC, while maintaining high consistency levels that is all the original samples are correctly classified by prototypes under the NN rule. Dasrathy's selection process is based on the concept of NUNS, the Nearest Unlike Neighbor Subset. With this concept, he first defines the initial subset as the entire training set, then for each instance in the subset the distance of the NUN is found and all instances with the same class whose distance is closer than the distance of the NUN are stored and counted as votes. The instance with the most votes is then designated as a prototype. All instances which contributed to the number of votes for this prototype is then removed from the other lists. The process is continued until the subset is no longer being reduced (**i need to clear this up**)

2.1.2 Instance Abstraction

Chang

Changs work deals with finding prototypes for nearest neighbor classifiers. The basic idea of his algorithm is to start with every sample in the training set as a prototype, then merge any two nearest prototype (p1 and p2 to form p*) with the same class as long as the recognition rate is not degraded. The new prototype p* can be formed by simply finding the average of p1 and p2 or the average

vector of weighted p1 and p2. p* will have the same class as the individual prototypes p1 and p2. The merging process will continue until the number of incorrect classifications of patterns in the data set starts to increase. Figure 2.1, shows Changs algorithm.

```
Step 1: Start with an arbitrary point tj
        in B* and assign it to A*
Step 2: For all points tk in B* such that
        class (tk) is not equal to class (tj),
        update bk to be the distance between
        tk ans tj if this daistance is smaller
        than the present bk. Otherwise, bk is
        unchanged.
Step 3: Amoung all the points in B*, find the
        point ts which has the smallest bs
        associated with it.
Step 4: If tj is not the nearest point ts such
        that the classes of tj and ts are
        different, go to Step 6. Othewise
        continue.
Step 5: Check whether or not d(tj, ts) is less
        than bj. If no, go to Step 6. If yes,
        let bj=d(tj, ts) and continue.
Step 6: Let j=s, move ts from B* to A*, and go
        to Step 2 until B* is empty. When B* is
        empty, the final b1,...,bm are the
        desired ones.
```

Figure 2.1: Chang's algorithm for finding prototypes

Chapter 3

CLIFF: Tool for Instance Selection

3.1 CLIFF: Tool for Instance Selection

CLIFF is a Prototype Learning tool based on a novel prototype learning algorithm which is defined by it's purpose - to reduce a training set via various selection and/or creation methods to produce *good* prototypes which increases the distance between NUNs and NLNs. These *good* prototypes are therefore a representation of the original training data set such that they maintain comparable or increased classification performance of a classifier.

CLIFF selects samples from a training set which best represents their respective target class. To accomplish this, a Bayesian ranking measure along with a support measure is used. First we assume that the target class is divided into one class as *best* and the other classes as *rest*. This makes it easy to find the attribute ranges of values which have a high probability of belonging to the current *best* class. The attribute ranges are found using an equal frequency binning algorithm which sorts attribute values into N equal frequency regions (making sure that there are no repeats in the numbers that are the boundaries between regions). The rank values are generated by applying Equation 3.1 to each attribute range. The remainder of this section further discusses how the attribute ranges and their corresponding ranks are generated.

```
Let likelihood = like
```

$$P(best|E) * support(best|E) = \frac{like(best|E)^2}{like(best|E) + like(rest|E)} \tag{3.1}$$

	Outlook	Temp(F)	Humidity	Windy?	Class
01.	Sunny	69	70	False	Lots
02.	Sunny	75	70	True	Lots
03.	Overcast	83	88	False	Lots
04.	Overcast	64	65	True	Lots
05.	Overcast	72	90	True	Lots
06.	Overcast	81	75	False	Lots
07.	Sunny	85	86	False	None
08.	Sunny	80	90	True	None
09.	Sunny	72	95	False	None
10.	Rain	65	70	True	None
11.	Rain	71	96	True	None
12.	Rain	70	96	False	Some
13.	Rain	68	80	False	Some
14.	Rain	75	80	False	Some

Figure 3.1: A log of some golf-playing behavior

The algorithm used in this work is a Support Based Bayesian Ranking (SBBR) algorithm. It's pseudo code is shown in Figure 3.2. The *best-rest* function divides the data set *D* into *best* and *rest* (for example, in Figure 3.1 if class *Lots* is selected as *best* then samples 1-6 would be in the best set while samples 7-14 would be in the rest set). As explained earlier the $EqualFrequencyBinning$ function in line 2 finds the attribute ranges for each attribute in the data set. For instance, looking at the attribute values for $Humidity$, if we want to get three(3) bins the following would be generated -

65 70 70 70 — 75 80 80 86 — 88 90 90 95 96 96.

Once the ranges for each attribute is found, lines 5-12 are applied to each range to produce a rank value. This rank value represents the probability a sample containing a value from a particular attribute range will be in the *best* class.

Finally to choose the samples which best represents a class, for each class a criteria is established where samples containing attribute range values with the highest *r* rank values of the highest *n* attributes are selected. For example, looking at the golf data set in Figure 3.1, if r=1 and n=4, the criteria for choosing the best samples in the *Lots* class could be [Overcast, 81-83, 75-88, False]. With this criteria, samples 3 and 6 would be selected. Please note that the attributes are ordered from most important to least important according to the highest rank value found in the respective attribute ranges. Therefore the highest *n* attributes is/are chosen according to this order.

```
DATA = [3, 5, 10, 20]
A = [Attributes]
BEST = [Instances with current best class]
REST = [Instances with other classes]
FREQ = [frequency]
LIKE = [likelihood]
EFB = [EqualFrequencyBinning]
BIN = [Values within Attribute range]
P_BEST = [Probability of BIN in BEST]
P_REST = [Probability of BIN in REST]

FOR EACH data IN DATA
 BIN_DATA = EFB on data
 FOR EACH attribute in A{
  FOR EACH BIN IN attribute{
   P_BEST = count(BEST) / count(data)
   P_REST = count(REST) / count(data)
   FREQ(BIN|BEST) = count(BIN in BEST) / count(BEST)
   FREQ(BIN|REST) = count(BIN in REST) / count(REST)
   LIKE(BEST|BIN) = FREQ(BIN|BEST) x P_BEST
   LIKE(REST|BIN) = FREQ(BIN|REST) x P_REST
   LIKE_BEST_REST = LIKE(BEST|BIN) + LIKE(REST|BIN)
   RANK = LIKE(BEST|BIN)^2 / LIKE_BEST_REST
   RETURN [BIN, RANK]
  END
 END
END
```

Figure 3.2: Pseudo code for Support Based Bayesian Ranking algorithm

Chapter 4

CLIFF Assessment

4.1 Data and Preprocessing Tools

4.1.1 Data Set Characteristics

Figure 4.1 lists the thirteen data sets used to assess CLIFF. The number of instances and attributes per instance are shown for each data set, along with the number of distinct classes of instances. All of these data sets were aquired from the UCI repository [16]. These data sets represent a variety of data types and characteristics. For example, three of the data sets (Sonar, Soybean and Splice) have large dimensions of 60, 35 and 60 respectively, while the others have dimensions in the range of four(4) to eighteen(18). Also the number of instances range from 148 (Lymph) to 3190 (Splice), while the number of classes range from two(2) to fifteen(15).

Assessing CLIFF on such a diverse set of data will give a good indication of the level of generalization CLIFF is capable off. However in the interest of speed, 2 pre-processing tools are applied to the high dimensional data sets in Figure 4.1. The following section details the algorithms used.

Data Set	Instances	Attributes	Class
Balance Scale	625	4	3
Breast Cancer	286	9	2
Heart (Cleveland)	297	13	5
Iris	150	4	3
Lymph	148	18	4
Pima Diabetes	768	8	2
Sonar	208	60	2
Soybean	562	35	15
Splice	3190	60	3
Vehicle	846	18	4
Vote	430	16	2

Figure 4.1: Data Set Characteristics

4.1.2 Pre-processing tools for Dimensionality Reduction

For this work, the pre-processing tools are aimed at dimensionality reduction using a novel feature subset selector (FSS) and FastMap [15]. Figure 4.2 shows which pre-processing tool is used on which data set if any at all while Figure 4.3 shows the pd and pf values of using various number of features for each data set. The highlighted f values are those choosen because of their high pds and low pfs, for example, when Soybean is FastMapped to f=16, it acheives the highest pd and lowest pf and so the data set with 16 features is used in future experiments in this work.

FastMap

The general goal of FASTMAP is to project items in a n dimensional to a d dimensional space, with $n > d$. The basis of each reduction is using the cosine law on the triangle formed by an object in the feature space and the two objects that are furthest apart in the current (pre-reduction) space (see Figure 4.4). These two objects are referred to as the pivot objects of that step in the reduction phase (n - d total pivot object sets). Finding the optimal solution of the problem of finding the two furthest apart points is a N^2 problem (where N is the total number of objects), but this is where the heuristic nature of FASTMAP comes into play.

Data Set	CLIFF	CLIFF + FSS	CLIFF + FastMap
Balance Scale	yes		
Breast Cancer	yes	yes	
Heart (Cleveland)	yes	yes	
Iris	yes		
Lymph	yes		
Pima Diabetes	yes	yes	
Sonar		yes	
Soybean			yes
Splice			yes
Vehicle		yes	
Vote		yes	

Figure 4.2: Data Set Characteristics

Instead of finding the absolute furthest apart points, FASTMAP takes a shortcut by first randomly selecting an object from the set, and then finding the object that is furthest from it and setting this object as the first pivot point. After the first pivot point is selected, FASTMAP finds the points farthest from this and uses it as the second pivot point. The line formed by these two points becomes the line that all of the other points will be mapped to in the new n - 1 dimension space.

FASTMAP uses the follow equation to calculate x_i, or the position of object O_i in the reduced space:

$$x_i = \frac{d_{a,i}^2 + d_{a,b}^2 - d_{b,i}^2}{2d_{a,b}} \tag{4.1}$$

This technique can be visualized by imagining the hyper-plane perpendicular to the line formed by pivot points, O_a and O_b, and projecting the new point onto this plane (Figure 4.5).

FASTMAP requires only $2D$ passes over D documents.

Feature Subset Selection (FSS)

Feature Subset Selection (FSS) is a technique that explores subsets of available features. There are many FSS algorithms out there including sequential forward selection and sequential backwards

Data Set	FSS	pd	pf%
Breast Cancer	f=2	53.2	46.8
	f=4	51.5	48.5
Heart (Cleveland)	f=2	0.0	9.4
	f=4	25.0	9.6
	f=8	25.0	10.0
Pima Diabetes	f=2	52.1	47.9
	f=4	59.5	40.5
Sonar	f=2	74.5	25.5
	f=4	78.0	22.0
	f=8	80.5	19.5
	f=16	78.6	21.4
Soybean	f=2	0.0	3.7
	f=4	50.0	1.9
	f=8	60.0	1.9
	f=16	66.7	1.2
Splice	f=2	24.3	14.8
	f=4	22.2	11.9
	f=8	22.0	13.8
	f=16	24.3	15.6
Vehicle	f=2	62.8	11.0
	f=4	66.3	10.2
	f=8	56.8	11.7
Vote	f=2	83.9	16.4
	f=4	85.7	14.3
	f=8	89.0	11.0

Figure 4.3: Choosing the best number of features for each data set. The best choice will have a high pd along with a low pf

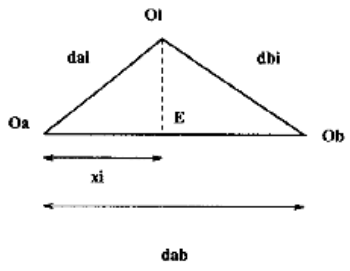

Figure 4.4: Example of using the cosine law to find the position of *Oi* in the dimension *k*

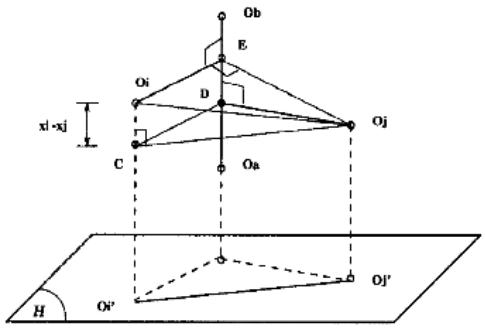

Figure 4.5: Projects of points O_i and O_j onto the hyper-plane perpendicular to the line O_aO_b

selection. However the simplest and fastest method involves ranking attribute from most important to least important. Infogain can be used effectively to rank attributes but in this work we use a novel ranking method for finding subsets for data set attributes. Using the ranking algorithm described in Section 3.1, first the highest rank of the attribute values is stored. The attributes are then sorted in the order of highest to lowest rank. With this information, the top n attributes are used in experiments.

4.2 CLIFF Assessment on Standard Data Sets

In this chapter, we evaluate CLIFF as a prototype learner on standard data sets in cross validation experiments. In the following sections we present results which show the probability of detection (pd) and probability of false alarm (pf) before and after the use of CLIFF.

4.2.1 Experimental Method

The effectiveness of CLIFF is measured using pd, pf and brittleness level (high, low) completed as folows: By allowing A, B, C and D to represent true negatives, false negatives, false positives and true positives respectfully, it then follows that *pd* also known as recall, is the result of true positives divided by the sum of false negative and true positives *D / (B + D)*. While pf is the result of: *C /*

(A + C). The *pd* and *pf* values range from 0 to 1. When there are no false alarms $pf = 0$ and at 100% detection, $pd = 1$.

The following sections describes the experiment and discusses the results.

4.2.2 Is CLIFF viable as a Prototype Learner for NNC?

The goal here is to see if the performance of CLIFF is comparable or better than the plain k nearest neighbor (KNN) algorithm. So in this experiment we compare the performance of predicting the target class using the entire training set to using only the prototypes generated by CLIFF. To accomplish this, our experiment design follows the pseudo code given in Figure 5.6 for the standard data sets. For each data set, tests were built from 20% of the data, selected at random. The models were learned from the remaining 80% of the data.

This procedure was repeated 5 times, randomizing the order of data in each project each time. In the end CLIFF is tested and trained 25 times for each data set.

```
DATA = [bc bs heart iris lym pima]
LEARNER = [KNN]
STAT_TEST = [Mann Whitney]

REPEAT 5 TIMES
 FOR EACH data IN DATA
  TRAIN = random 90% of data
  TEST = data - TRAIN

  \\Construct model from TRAIN data
  MODEL = Train LEARNER with TRAIN
  \\Evaluate model on test data
  [pd, pf] = MODEL on TEST
 END
END
```

Figure 4.6: Pseudo code for Experiment

Results from Experiment

Results for Experiment 1 for the 4 data sets distinguished by the number of clusters. Here for the upper and lower tables f=4 is used while r=1 is used for the upper table and r=2 for the lower table.

ir	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	91	100	100	
	AFTER	73	100	100	
pf	BEFORE	0	0	5	
	AFTER	0	0	8	
					0 50 100

bc	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	29	47	85	
	AFTER	23	40	95	
pf	BEFORE	11	21	69	
	AFTER	5	12	73	
					0 50 100

bs	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	0	68	79	
	AFTER	44	53	61	
pf	BEFORE	3	18	25	
	AFTER	12	20	29	
					0 50 100

ht	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	8	21	50	
	AFTER	0	0	50	
pf	BEFORE	6	11	20	
	AFTER	0	4	20	
					0 50 100

ly	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	0	71	88	
	AFTER	0	69	100	
pf	BEFORE	0	0	24	
	AFTER	0	0	25	
					0 50 100

pm	Treatment	25%	50%	75%	Q1 median Q3
pd	BEFORE	40	58	72	
	AFTER	24	46	87	
pf	BEFORE	28	36	60	
	AFTER	13	31	76	
					0 50 100

Figure 4.7: Probability of Detection (PD) and Probability of False Alarm (PF)results

group	pd (Before $\rightarrow After$)	pf (Before $\rightarrow After$)	position (Before $\rightarrow After$)	data	$\lvert data \rvert$
a	same	same	same	pm	1
b	same	same	increase	bc ir ly	3
c	decrease	decrease	increase	ht	1
d	decrease	increase	increase	bs	1

Figure 4.8: Summary of Mann Whitney U test results (95% confidence): moving from Befroe to After.

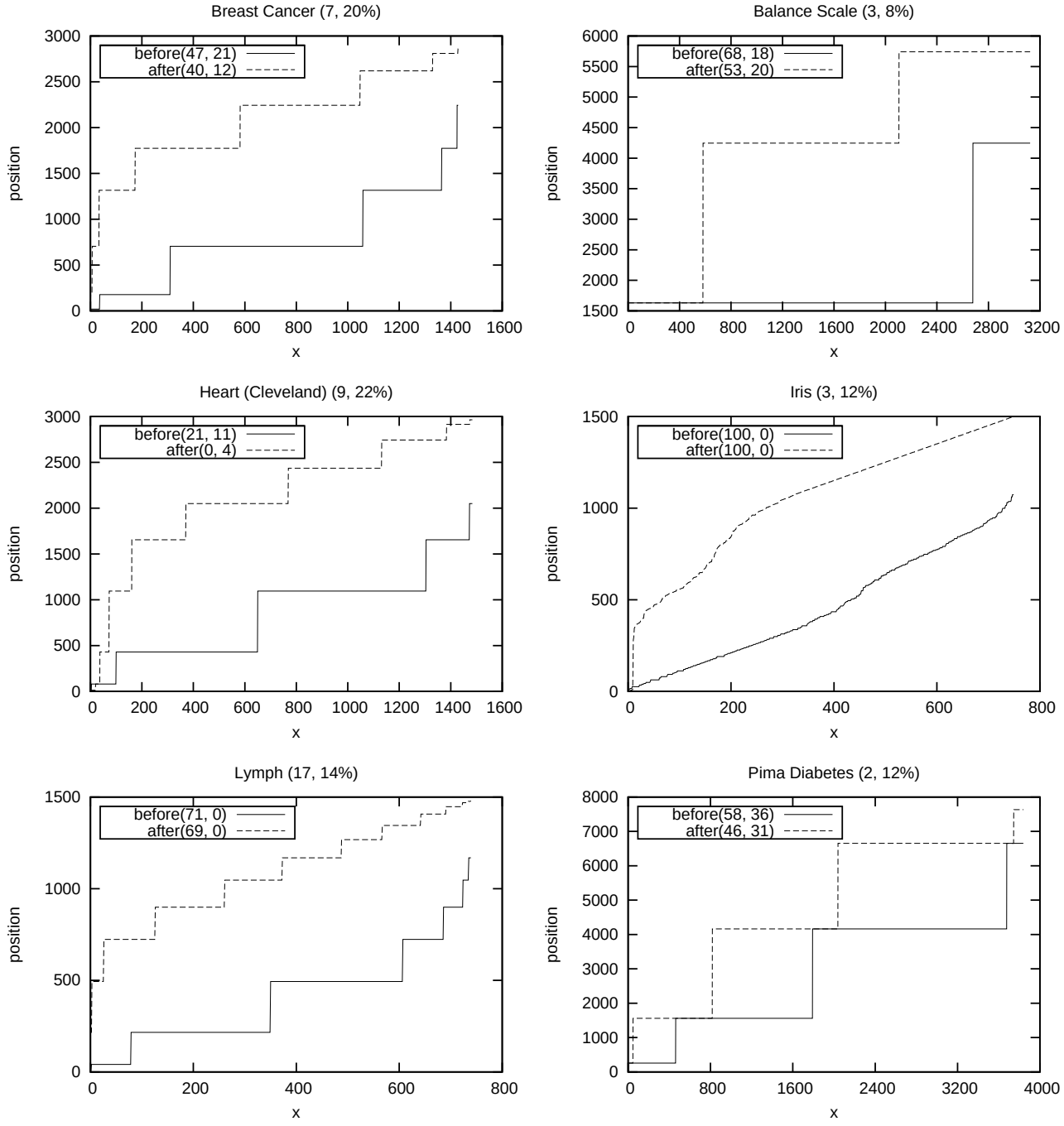

Figure 4.9: Position of values in the 'before' and 'after' population with data set at 3, 5, 10 and 20 clusters. The first row shows the results for r=1 while the second row shows the results for r=2

Chapter 5

Case Study: Solving the Problem of Brittleness in Forensic Models

5.1 Introduction

The principal goal of forensic evaluation models is to check that evidence found at a crime scene is (dis)similar to evidence found on a suspect. In creating these models, attention is given to the significance level of the solution however the *brittleness* level is never considered. The *brittleness* level is a measure of whether a solution comes from a region of similar solutions or from a region of dissimilar solutions. We contend that a solution coming from a region with a low level of brittleness i.e. a region of similar solutions, is much better that one from a high level of brittleness - a region of dissimilar solutions.

The concept of *brittleness* is not a stranger to the world of forensic science, in fact it is recognized as the "fall-off-the-cliff-effect", a term coined by Ken Smalldon. In other words, Smalldon recognized that tiny changes in input data could lead to a massive change in the output. Although Walsh [35] worked on reducing the brittleness in his model, to the best of our knowledge, no work been done to quantify brittleness in current forensic models or to recognize and eliminate the

causes of brittleness in these models.

In our studies of forensic models for evaluation particularly in the sub-field of glass forensics, we conjecture that brittleness is caused by the following:

1. A tiny error(s) in the collection of data;

2. Inappropriate statistical assumptions, such as assuming that the distributions of refractive indices of glass collected at a crime scene or a suspect obeys the properties of a normal distribution;

3. and the use of measured parameters from surveys to calculate the *frequency of occurrence* of trace evidence in a population

In this work we quickly eliminate the two(2) latter causes of brittleness by using simple classification methods such as k-nearest neighbor (KNN) which are neither concerned with the distribution of data nor the frequency of occurrence of the data in a population. To reduce the effects of errors in data collection, a novel prototype learning algorithm (PLA) is used to augment KNN. Basically this PLA selects samples from the data which best represents the region or neighborhood it comes from. In other words, we expect that samples which contain errors would be poor representatives and would therefore be eliminated from further analysis. This leads to neighbourhoods with different outcomes being futher apart from each other.

In the end our goal for this work is threefold. First we want to develop a new generation of forensic models which avoids inappropriate statistical assumptions. Second, the new models must not be *brittle*, so that they do not change their interpretation without sufficient evidence and third, provide not only an interpretation of the evidence but also a measure of how reliable the interpretation is, in other words, what is the brittleness level of the model.

Our research is guided by the following research questions:

- Using KNN as a model, what is the best K for each data set?

- Are the results of using KNN better or comparable to current models which use statistical assumptions and surveys
- Does prototype learning reduce brittleness?
- Do the results of applying a PLA differ significantly from results of not applying a PLA?

5.2 Visualization of Brittleness

This work is motivated by a recent National Academy of Sciences report titled "Strengthening Forensic Science" [32]. This report took special notice of forensic interpretation models stating:

> With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source. [32], p6

In this study we visualize the inconsistencies of four(4) of these forensic methods in one way. By simply plotting the measurements derived from evidence from a crime scene denoted as x and suspect (y), against the results of interpretation.

The rest of this section gives details of the four(4) forensic models evaluated in this work, followed by the visualization of these models to highlight their brittleness which results in the inconsistencies in model results.

5.3 Glass Forensic Models

This section provides an overview of the following glass forensic models used in this work to show brittleness.

1. The 1978 Seheult model [33]

2. The 1980 Grove model [19]

3. The 1995 Evett model [14]

4. The 1996 Walsh model [35]

5.3.1 Seheult 1978

Seheult [33], examines and simplifies Lindley's [31] 6th equation for real-world application of Refractive Index (RI) Analysis. According to Seheult:

> A measurement x, with normal error having known standard deviation σ, is made on the unknown refractive index Θ_1 of the glass at the scene of the crime. Another measurement y, made on the glass found on the suspect, is also assumed to be normal but with mean Θ_2 and the same standard deviation as x. The refractive indices Θ are assumed to be normally distributed with known mean μ and known standard deviation τ. If I is the event that the two pieces of glass come from the same source($\Theta_1 = \Theta_2$) and $\bar{I}$ the contrary event, Lindley suggests that the odds on identity should be multiplied by the factor
>
> $$\frac{p(x,y|I)}{p(x,y|\bar{I})} \tag{5.1}$$
>
> In this special case, it follows from Lindley's 6th equation that the factor is
>
> $$\frac{1+\lambda^2}{\lambda(2+\lambda^2)^{1/2}} \quad ^{-\frac{1}{2(1+\lambda^2)}\cdot(u^2-v^2)} \tag{5.2}$$
>
> Where
>
> $$\lambda = \frac{\sigma}{\tau}, u = \frac{x-y}{\sigma\sqrt{2}}, v = \frac{z-\mu}{\tau(1+\frac{1}{2}\lambda^2)^{\frac{1}{2}}}, z = \frac{1}{2}(x+y)$$

5.3.2 Grove 1980

By adopting a model used by Lindley and Seheult, Grove proposed a non-Bayesian approach based on likelihood ratios to solve the forensic problem. The problem of deciding whether the fragments have come from common source is distinguished from the problem of deciding the guilt or innocence of the suspect. To explain his method, Grove first reviewed Lindley's method. He argued that we should, where possible, avoid parametric assumptions about the underlying distributions. Hence, in discussing the respective roles of θ_1 and θ_2 Grove did not attribute any probability distribution to an unknown parameter without special justification. So when considering (θ_1 != θ_2), $\bar{I}$ can be interpreted as saying that the fragments are present by chance entailing a random choice of value for θ_2. The simplified likelihood ratio obtained from the Grove's derivation is:

$$\frac{\tau}{\sigma} \cdot e^{\left\{ \frac{-(X-Y)^2}{4\sigma^2} + \frac{(Y-\mu^2)}{2\tau^2} \right\}} \tag{5.3}$$

We are of course only concerned with the evidence about I and $\bar{I}$ so far as it has the bearing on the guilt or innocence of the suspect. Grove also considered the Event of Guilty factor G in the calculation of likelihood ratio (LR). Therefore the LR now becomes

$$p(X,Y|G)/p(X,Y|\bar{G}) \tag{5.4}$$

Here in the expansion event T, that fragments were transferred from the broken window to the suspect and persisted until discovery and event A,that the suspect came into contact with glass from other source. Here p(A/G)=p(A/$\bar{G}$)=Pa and p(T/G)= Pt. The resulting expression is

$$\frac{P(X,Y,S|G)}{P(X,Y,S|\bar{G})} = 1 + Pt\left\{ (\frac{1}{Pa} - 1)\frac{p(X,Y|I)}{p(X,Y|\bar{I})} - 1 \right\} \tag{5.5}$$

5.3.3 Evett 1995

Evett et al used data from forensic surveys to create a Bayesian approach in determining the statistical significance of finding glass fragments and groups of glass fragments on individuals associated with a crime [14].

Evett proposes that likelihood ratios are well suited for explaining the existence of glass fragments on a person suspected of a crime. A likelihood ratio is defined in the context of this paper as the ratio of the probability that the suspected person is guilty given the existing evidence to the probability that the suspected person is innocent given the existing evidence. The given evidence, as it applies to Evett's approach, includes the number of different sets of glass and the number of fragments in each unique group of glass.

The Lambert, Satterthwaite and Harrison (LSH) survey used empirical evidence to supply probabilities relevant to Evett's proposal. The LSH survey inspected individuals and collected glass fragments from each of them. These fragments were placed into groups based on their refractive index (RI) and other distinguishing physical properties. The number of fragments and the number of sets of fragments were recorded, and the discrete probabilities were published. In particular, there are two unique probabilities that are of great interest in calculating Evett's proposed likelihood ratio.

- S, the probability of finding N glass *fragments* per group
- P, the probability of finding M *groups* on an individual.

The following symbols are used by Evett to express his equations:

- P_n is the probability of finding n groups of glass on the surface of a person's clothes
- T_n is the probability that n fragments of glass would be transferred, retained and found on the suspect's clothing if he had smashed the scene window

- S_n is the probability that a group of glass fragments on a person's clothing consists of n fragments
- f is the probability that a group of fragments on person's clothing would match the control sample
- λ is the expected number of glass fragments remaining at a time, t

Evett utilizes the following equations to determine the likelihood ratio for the first case described in his 1994 paper. In this case, a single window is broken, and a single group of glass fragments is expected to be recovered.

$$LR = \frac{P_0 T_n}{P_1 S_n f} + T_0 \tag{5.6}$$

$$T_n = \frac{e^{-\lambda}\lambda^n}{n!} \tag{5.7}$$

5.3.4 Walsh 1996

The equation presented by Walsh [35] is similar to one of Evett's. The difference is that Walsh argues that instead of incorporating grouping and matching, only grouping should be included. Walsh says this is because match/non-match is really just an arbitrary line. He examines the use of a technique in interpreting glass evidence of a specific case. This technique is as follows:

$$\frac{T_L P_0 p(\bar{X}, \bar{Y} | S_y, S_x)}{P_1 S_L f_1} \tag{5.8}$$

Where

- T_L = the probability of 3 or more glass fragments being transferred from the crime scene to the person.

- P_0 = the probability of a person having no glass on their clothing
- P_1 = the probability of a person having one group of glass on their clothing
- S_L = the probability that a group of glass on clothing is 3 or more fragments
- $\bar{X}$ and $\bar{Y}$ are the mean of the control and recovered groups respectively
- S_x and S_y are the sample standard deviations of the control and recovered groups respectively
- f_1 is the value of the probability density for glass at the mean of the recovered sample
- $p(\bar{X}, \bar{Y}|S_y, S_x)$ is the value of the probability density for the difference between the sample means

5.4 Visualization of Brittleness in Models

The result of applying the visualization technique i.e. plotting the measurements derived from evidence from a crime scene denoted as x and suspect (y), against the results of interpretation on the glass forensic models are shown in Figure 5.1.

For the first two(2) models the x and y axes represent the mean refractive index (RI) values of evidence from a crime scene and suspect respectively. While the x axis of the Walsh model represents $f1$ is the value of the probability density for glass at the mean of the recovered sample and the y axis represents the value of the probability density for the difference between the sample means. The x and y axes of the Evett model represents λ and $f - values$ respectively. The z axis of all the models represent the likelihood ratio (LR) generated from these models, in other words, the significance of the match/non-match of evidence to an individual or source.

Using data donated by the Royal Canadian Mounted Police (RCMP), values such as the RI ranges and their mean, were extracted to generate random samples for the forensic glass models. In all four(4) models 1000 samples are randomly generated for the variables in each model. For

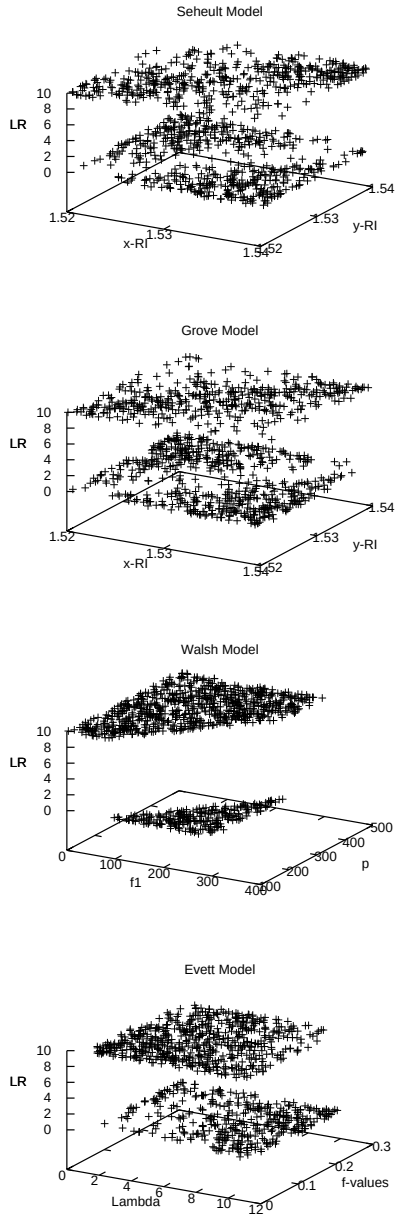

Figure 5.1: Visualization of four(4) glass forensic models

instance, in the Seheult model, each sample looks like this: [x, y, σ, μ, τ]. The symbols are explained in 5.3.1.

In Figure 5.1 - the sehult and grove models, brittleness or Smalldon's "fall-off-the-cliff-effect"

is clearly demonstrated. These models proposed by Seheult (Section 5.3.1) and Grove (5.3.2) respectively, show how the likelihood ratio changes (on the vertical axis) as we try different values from the refractive index of from glass from two sources (x and y). This model could lead to incorrect interpretations if minor errors are made when measuring the refractive index of glass samples taken from a suspect's clothes. Note how, near the region where x=y, how tiny changes in the x or y refractive indexes can lead to dramatic changes in the likelihood ratio (from zero to one).

The visualization of the Evett (5.3.3) and Walsh (5.3.4) models show similar brittleness when the likelihood ratios are 0 and 1. For Walsh, values located at the edge of a cliff a LR=1 can easily become LR=0 at the smallest change in the $f1$ or p values. While Evett will cause problems because a small change occurs with any sample it is possible for the LR to change.

From these visualizations it is obvious that the concern of the National Academy of Sciences report [32] mentioned earlier in this section is a valid one. So how can this concern be alleviated? We propose not only including a *brittleness* measure to a forensic method as a solution, but also moving away from forensic models which use a Bayesian approach [12–14, 33, 35], and statitical assumptions [19, 33, 35].

The following sections gives details of the models used in this work as well as the data set used to evaluate the models.

5.5 Introduction

If standard methods are brittle what can we do? We seek our answer to this question in the work of [23], and we explore an intuition that to reduce brittleness, data with dissimilar outcomes should not be close neighbors. In this section the details of CLIFF's core procedure and tools are discussed. Included in this discussion is a sub-section which further explores our intuition for brittleness reduction and our tool borne from this intuition - the CLIFF selector.

The Design of CLIFF is deeply rooted in the work of [23]. In their work, analysis is done using

Chemometrics, an application of mathematical, statistical and/or computer science techniques to chemistry. In the work done by [23], Chemometrics using computer science techniques is applied to analyze the infrared spectra of the clear coat layer of a range of cars. The analysis proceeded as follows:

- Agglomerative hierarchical clustering (AHC) for grouping the data into classes
- Principal component analysis (PCA) for reducing dimensions of the data
- Discriminant analysis for classification i.e. associating an unknown sample to a group or region

This technique produced a strong model which achieved 100% accuracy i.e. when validated by removing random samples from the model, all the samples were correctly assigned. The goal of CLIFF is not only to create a strong forensic model but also to show how strong the model is. To achieve this CLIFF includes a brittleness measure as well as a method to reduce brittleness. Also, in an effort to keep CLIFF simple, we substituted different tools to preform the analysis done in [23]. For instance Kmeans is used instead of AHC for grouping the data into classes. FastMap is used for dimensionality reduction and K-nearest neighbor is used for classification. The basic operation of CLIFF is shown in Figure 5.2. The data is collected and the dimensions is reduced if necessary. Clusters are then created from the data and classification is done along with a brittleness measure (further discussed in Section 5.9). Finally, we test if brittleness can be reduced using a novel prototype learning technique (Instance Selection).

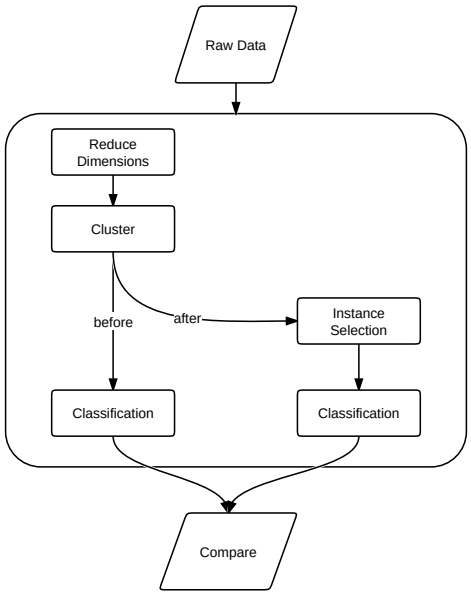

Figure 5.2: Proposed procedure for the forensic evaluation of data

5.6 Dimensionality Reduction

5.6.1 Principal Component Analysis

The goal of Principal component analysis (PCA) is to reduce the number of variables or dimensions of a data set which has a large number of correlated variables while maintaining as much of the data variation as is possible. The result of this serves two main purposes:

1. To simplify analysis and

2. To aid in the visualization of the data

To achieve this goal, the data set is transformed to a new set of variables which are not correlated and which are ordered so that the first few principal components (PCs) retain most of the variation present in all of the original variables [22]. Let us look at an example. Figure 5.3 shows a visualization of Fisher's five-dimensional iris data on a two-dimensional scatter plot. First,

PCs are extracted from the four continuous variables (sepal-width, sepal-length, petal-width, and petal-length). Second, these variables are projected onto the subspace formed by the first two components extracted. Finally this two-dimensional data is shown on a scatter-plot in Figure 5.3. The fifth dimension (species) is represented by the color of the points on the scatter-plot.

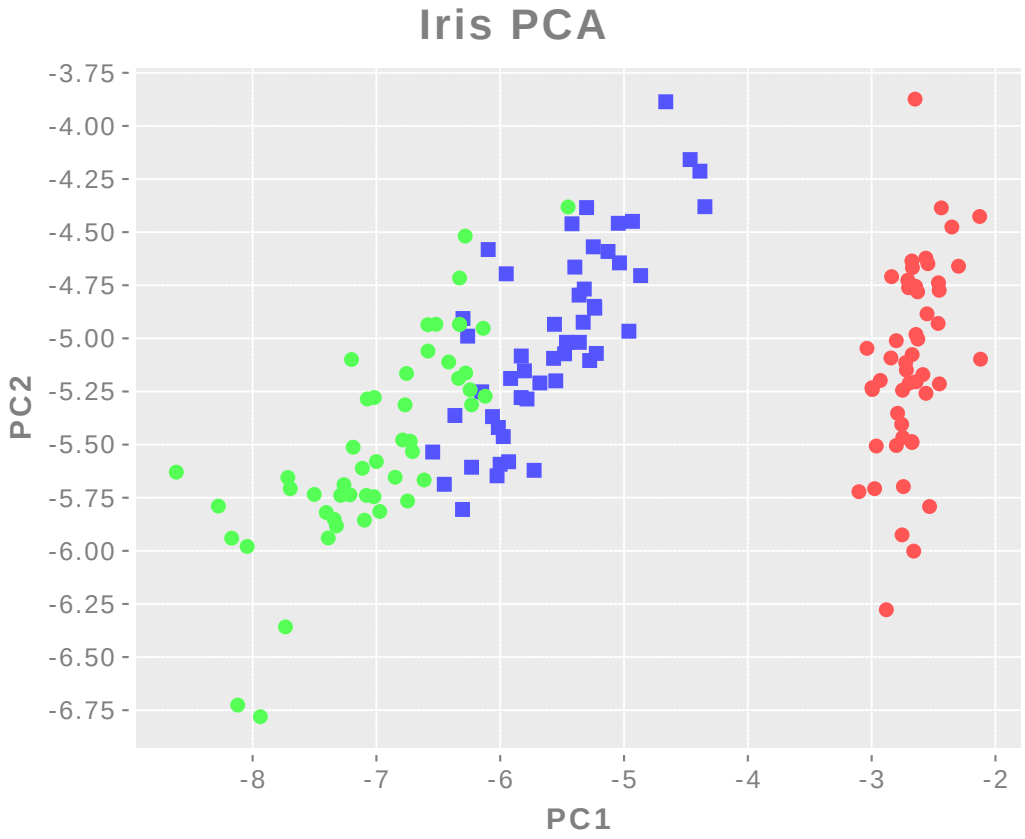

Figure 5.3: PCA for iris data set

The data used in our experiments contains 1151 attributes and 185 instances. Using the data set as is would cause us to create a model that is computationally expensive and likely to produce unacceptable results such as a high false positive values caused by redundant and noisy data. To avoid this foreseen problem, we turn to dimensionality reduction.

Dimensionality Reduction refers to reducing high-dimensional data to low dimensional data. This is accomplished by attempting to summarize the data by using less terms than needed. While

this reduces the overall information available and thus a level of precision, it allows for easy visualization of data otherwise impossible to visualize. Some algorithms that can be used for Dimensionality Reduction are Principle Component Analysis (PCA), and FastMap.

The data used in this work contains 1,151 variables and 185 samples. To perform an analysis on this data set we must first reduce the number of variables used. In [23], PCA is used to perform dimensionality reduction. PCA can be defined as "the orthogonal projection of the data onto a lower dimensional linear space". In other words, looking at our data set, our goal is to project the data onto a space having dimensionality that is less than 1,151 (M ¡ 1,151) while maximizing the variance of the projected data [5]. In [23], two techniques - Pearson correlation and covariance for comparison of the two, were used to determine an appropriate value for M (M = 4).

To speed things up a little, in our model we use *FastMap* to reduce the dimensions of the data set. In FastMap the basis of each reduction is using the cosine law on the triangle formed by an object in the feature space and the two objects that are furthest apart in the current (pre-reduction) space. These two objects are referred to as the pivot objects of that step in the reduction phase (M total pivot object sets). Finding the optimal solution of the problem of finding the two furthest apart points is an N squared problem (where N is the total number of objects), but this is where the heuristic nature of FastMap comes into play. Instead of finding the absolute furthest apart points, FastMap takes a shortcut by first randomly selecting an object from the set, and then finding the object that is furthest from it and setting this object as the first pivot point. After the first pivot point is selected, FastMap finds the points farthest from this and uses it as the second pivot point. The line formed by these two points becomes the line that all of the other points will be mapped to in the new M dimension space. (Further details of this algorithm can be found elsewhere [15]).

To determine the appropriate value for M using FastMap, we experimented experimented with different values for M. Figure 5.4 shows results for various K-nearest neighbor classifiers (discussed further in Sections 5.8 and 4), with M fixed at 2, 4, 8 and 16. When M is 2 or 4 100% of the validation samples are predicted correctly (pd) and 0% are predicted incorrectly (pd). For this

reason, our model model is analysed using M = 4.

5.7 Clustering

Clustering is the second step in the CLIFF tool and can be defined as the grouping of the samples into groups whose members are similar in some way. The samples that belong to two different clusters are dissimilar. The major goal of clustering is to determine the intrinsic grouping in the set of unlabelled data. In most of the clustering techniques, distance is the major criteria. Two objects are similar if they are close according to the given distance.

CLIFF clusters using K-means. The Figure 5.5 represents the pseudo code for the K-means algorithm. The idea behind K-means clustering is done by assuming some arbitrary number of centroids, then the objects are associated to nearest centroids. The centroids are then moved to center of the clusters. These steps are repeated until a suitable level of convergence is attained.

5.8 Classification with KNN

K-nearest neighbor (KNN) classification is a simple classification method usually used when there is little or no prior knowledge about the distribution of data. KNN is described in [11] as follows: Stores the complete training data. New samples are classified by choosing the majority class among the k closest examples in the training data. For our particular problem, we used the Euclidean, i.e. sum of squares, distance to measure the distance between samples. Finally, to determine a value for k, we investigated the performance of six (6) KNN classifiers where k is fixed at 2, 4, 8 and 16. Figure 5.4 shows the results which indicate that using KNN classifiers where k is equal to 4, 8 or 16, the validation of samples is 100%. For CLIFF k = 4 is used.

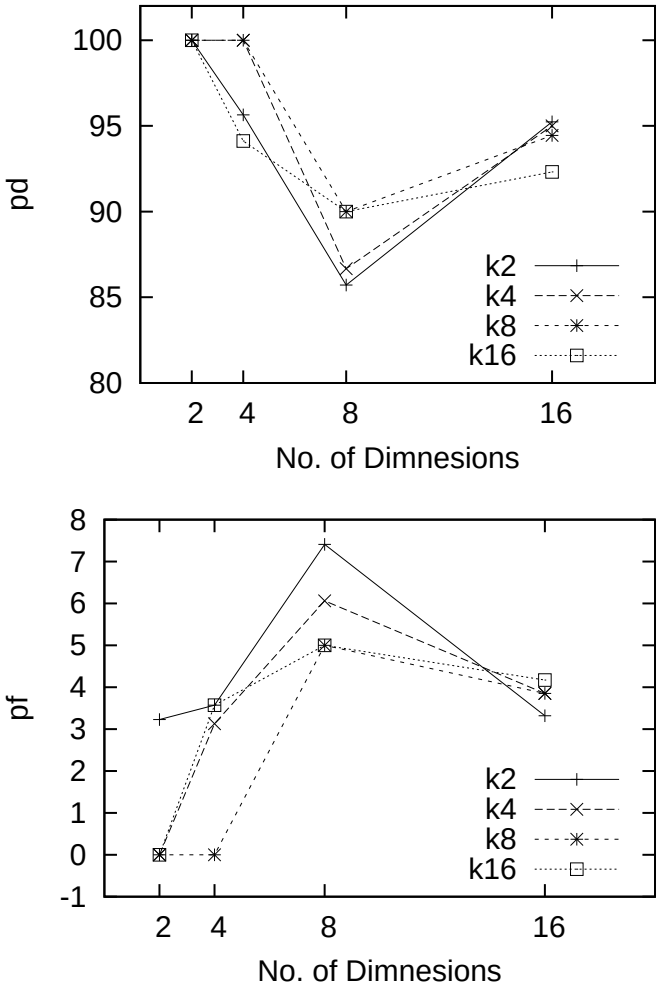

Figure 5.4: Probability of detection (pd) and Probability of False alarms (pf) using fixed values for dimensions and fixed k values for k-nearest neighbor

```
DATA = [3, 5, 10, 20]
k = [1, ..., Number of clusters]
STOP = [Stopping criteria]

FOR EACH data IN DATA
 N = count(data)
 WHILE STOP IS FALSE
  // Calculate membership in clusters
  FOR EACH data point X IN data
   FIND NEAREST CENTROID_k
   ADD TO CLUSTER_k
  END

  // Recompute the centroids
  FOR EACH CLUSTER
   FIND NEW CENTROIDS
  END

  // Check stopping criteria
  [TRUE or FALSE] = STOP
 END
END
```

Figure 5.5: Pseudo code for K-means

5.9 The Brittleness Measure

Calculating the brittleness measure is a novel operation of CLIFF. We use the brittleness measure in this work to determine if the results of CLIFF comes from a region where all the possible results are (dis)similar. For the purpose of this work the optimal result will come from a region of similar results. To make this determination, using each sample from a validation set, once each sample from this set has been classified, the distance from the nearest unlike neighbor (NUN) i.e. the distance from a sample with a different class and the distance from the nearest like neighbor (NLN) i.e. the distance from a sample with the same class is recorded. Recall that brittleness is a small change can result in a different outcome, so here the closer the distances of NUN to NLN

the more brittle the model. So an ideal result will have the greatest distance between NUNs and NLNs.

The brittleness measure will give an output of either *high* or *low*: high indicating that there is no significant difference between the NUN and NLN values, while *low* indicates the opposite. The significance of these values was calculated using the Mann-Whitney U test. This is a non-parametric test which replaces the distance values with their rank or position inside the population of all sorted values.

Equation 5.9 embodies our definition of brittleness: if the significance of NUN values are less than or equal to the NLN values, then an unacceptable level of brittleness is present in the model.

$$[NUN <= NLN] ==> BRITTLENESS \tag{5.9}$$

In this chapter, we evaluate CLIFF as a forensic model on a data set donated by [23] in cross validation experiments. First, we describe the data set and experimental procedures. Next we present results which show the probability of detection (pd), probability of false alarm (pf) and brittleness level of CLIFF before and after the use of the selector.

5.10 Data Set and Experimental Method

The data set used in this work is donated by [23]. It contains 37 samples each with five(5) replicates (37 x 5 = 185 instances). Each instance has 1151 infrared measurements ranging from 1800-650cm-1. (Further details of this algorithm can be found elsewhere [23]). For our experiments we took the original data set and created four (4) data sets each with a different number of clusters (3, 5, 10 and 20) or groups. These clusters were created using the K-means algorithm (Figure 5.5).

The effectiveness of CLIFF is measured using pd, pf and brittleness level (high, low) completed as folows: By allowing A, B, C and D to represent true negatives, false negatives, false positives and true positives respectfully, it then follows that *pd* also known as recall, is the result of true

positives divided by the sum of false negative and true positives *D / (B + D)*. While pf is the result of: *C / (A + C)*. The pd and pf values range from 0 to 1. When there are no false alarms $pf = 0$ and at 100% detection, $pd = 1$.

The brittleness level measure is conducted as follows: First we calculate Euclidean distances between the validation or testing set which has already been validated and the training set. For each instance in the validation set the distance from its nearest like neighbor (NLN) and its nearest unlike neighbor (NUN) is found. Using these NLN and NUN distances from the entire validation set a Mann-Whitney U test was used to test for statistical difference between the NLN and NUN distances. The following sections describes two experiments and discusses their results.

5.11 Experiment 1: KNN as a forensic model?

Our goal is to determine if KNN is an adequate model for forensic evaluation. In other words, can it be used in preference to current statistical models? To answer this question, our experiment design follows the pseudo code given in Figure 5.6 for the four (4) data sets created from the original data set. For each data set, tests were built from 20% of the data, selected at random. The models were learned from the remaining 80% of the data.

This procedure was repeated 5 times, randomizing the order of data in each project each time. In the end CLIFF is tested and trained 25 times for each data set.

5.11.1 Results from Experiment 1

Figure 5.7 shows the 25%, 50% and 100% percentile values of the pd, pf and position values in each data set when r=1 (upper table) and r=2 (lower table. Next to these is the brittleness signal where *high* signals an unacceptable level of brittleness and *low* signals an acceptable level of brittleness. The results show that the brittleness level for each data set is *low*. The pd and pf results are promising showing that 50% of the pd values are at or above 95% for the data set with

```
DATA = [3, 5, 10, 20]
LEARNER = [KNN]
STAT_TEST = [Mann Whitney]

REPEAT 5 TIMES
 FOR EACH data IN DATA
  TRAIN = random 90% of data
  TEST = data - TRAIN

  \\Construct model from TRAIN data
  MODEL = Train LEARNER with TRAIN
  \\Evaluate model on test data
  [brittleness] = STAT_TEST on NLN and NUN
  [pd, pf, brittleness] = MODEL on TEST
 END
END
```

Figure 5.6: Pseudo code for Experiment 1

3 clusters and at 100% for the other data sets. While 50% of the pf values are at 3% for 3 clusters and 0% for the others. These results show that our model is highly discriminating and can be used successfully in the evaluation of trace evidence.

5.12 Experiment 2: Can brittleness be reduced?

The first experiment shows that KNN creates strong models for forensic evaluation, with high pd's, low pf's and low brittleness levels. With experiment 2 we want to find out if these results can be improved by reducing brittleness further. Since we believe that it is the nearness of unlike neighbors which causes the brittleness (See Equation 5.9), in this section we evaluate the CLIFF selector which selects a subset of data from each cluster which best represents the cluster in hopes that this increases the distance between like neighbors and therefore decrease brittleness while maintaining comparable pd and pf results from experiment 1. Also we expect that the position

		Before percentiles			
Clusters	Types	25%	50%	75%	Brittlness Level
3	pd	90	95	100	low
	pf	0	3	4	
	position	264	614	1068	
5	pd	94	100	100	low
	pf	0	0	0	
	position	374	855	1225	
10	pd	75	100	100	low
	pf	0	0	0	
	position	361	783	1254	
20	pd	0	100	100	low
	pf	0	0	3	
	position	377	762	1256	

		Before percentiles			
Clusters	Types	25%	50%	75%	Brittlness Level
3	pd	89	94	100	low
	pf	0	0	4	
	position	419	905	1351	
5	pd	94	100	100	low
	pf	0	0	0	
	position	437	903	1297	
10	pd	50	100	100	low
	pf	0	0	3	
	position	442	908	1354	
20	pd	0	67	100	low
	pf	0	0	0	
	position	437	896	1345	

Figure 5.7: Results for Experiment 1 for the 4 data sets distinguished by the number of clusters. Here for the upper and lower tables n=4 is used while r=1 is used for the upper table and r=2 for the lower table.

values well be greater than those in the experiment 1.

The design for this experiment can be seen in Figure 5.8. It is similar to that in Figure 5.6, however, the CLIFF selector is included and is described in 3.1.

```
DATA = [3, 5, 10, 20]
LEARNER = [KNN]
STAT_TEST = [Mann Whitney]
SELECTOR = [CLIFF selector]

REPEAT 5 TIMES
 FOR EACH data IN DATA
  TRAIN = random 90% of data
  TEST = data - TRAIN

  \\CLIFF selector: select best from clusters
  N_TRAIN = SELECTOR with TRAIN

  \\Construct model from TRAIN data
  MODEL = Train LEARNER with N_TRAIN
  \\Evaluate model on test data
  [brittleness] = STAT_TEST on NLN and NUN
  [pd, pf, brittleness] = MODEL on TEST
 END
END
```

Figure 5.8: Pseudo code for Experiment 2

5.12.1 Results from Experiment 2

Figure 5.9 shows results for 5 and 10 clusters remain the same for 50% of the pd and pf values while for 3 and 20 clusters the pd's have decreased to 82% and 67% respectively. Also the brittleness level remains low for each data set. The results shown in Figure 5.9 does not provide any information about the difference between the low level of brittleness between Figure 5.7 and Figure 5.9, however the model remains strong. Figure 5.10 illustrates the reduction of brittleness after

the CLIFF selector is applied. Mann Whitney U test was also applied to these results to see if there was a statistical difference between the before and after results. The test indicated that the $after$ results are better than $before$ (see Figure 5.11). So brittleness can be reduced while maintaining comparable results.

In summary, by using CLIFF, inappropriate statistical assumptions about the data are avoided. We found a successful way to reduce any brittleness found, to create strong forensic evaluation models. One important point to note here also is this: In order to evaluate data sets with multiple variables, a host of new statistical models has been built [1,2,27,28,36,37]. This has been the case with forensic scientists building these models for glass interpretation when using the elemental composition of glass rather than just the refractive indices. On the other hand, with CLIFF an increase in the number of variables used does not signal the need to create a new model, it works with any data set.

Clusters	Types	After percentiles 25%	50%	75%	Brittleness Level
3	pd	49	82	100	low
	pf	0	9	20	
	position	787	1228	1609	
5	pd	94	100	100	low
	pf	0	0	0	
	position	563	988	1532	
10	pd	60	100	100	low
	pf	0	0	3	
	position	578	1048	1463	
20	pd	0	67	100	low
	pf	0	0	3	
	position	601	1081	1481	

Clusters	Types	After percentiles 25%	50%	75%	Brittleness Level
3	pd	89	100	100	low
	pf	0	0	5	
	position	633	1047	1432	
5	pd	90	100	100	low
	pf	0	0	0	
	position	507	982	1465	
10	pd	100	100	100	low
	pf	0	0	0	
	position	506	968	1426	
20	pd	0	80	100	low
	pf	0	0	0	
	position	495	957	1424	

Figure 5.9: Results for Experiment 2 for the 4 data sets distinguished by the number of clusters. Here for the upper and lower tables n=4 is used while r=1 is used for the upper table and r=2 for the lower table.

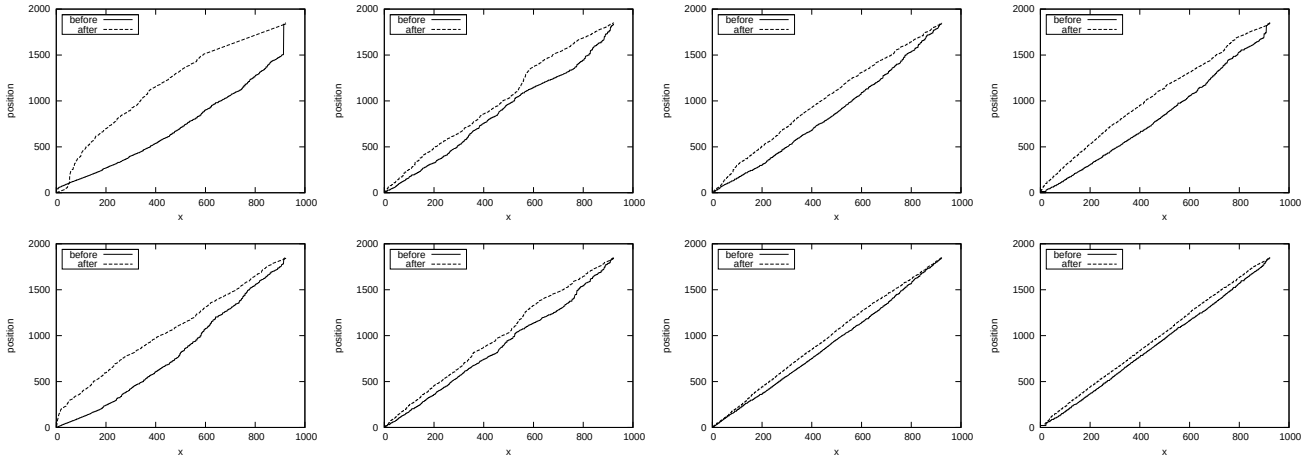

Figure 5.10: Position of values in the 'before' and 'after' population with data set at 3, 5, 10 and 20 clusters. The first row shows the results for r=1 while the second row shows the results for r=2

Clusters	Treatments	Significance
3	before after	-1
5	before after	-1
10	before after	-1
20	before after	-1

Figure 5.11: Results for Experiment 2 of before and after results. -1 indicates that the after is better than before

Chapter 6

Conclusion

The principal purpose of this work was to address the concern of the National Academy of Sciences which stated in a report [32] that:

> With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source. [32], p6

Our answer to this is, a novel approach for the evaluation of trace forensic evidence. With a data set donated by [23], made up of the infrared spectra of the clear coat layer of a range of cars, we showed that:

- CLIFF creates strong models with *low* brittleness levels
- The CLIFF selector, based on a PLA, can further reduce the brittleness of a model
- The levels of brittleness differ significantly before and after the use of the CLIFF selector (Mann Whitney U test)

It is our intent that this work open the eyes of the forensic scientist to the real problem of *brittleness* which exists in current forensic models. We hope in the future that the scientist, when verifying a model, they include a brittleness measure along with their evaluation of forensic evidence as done in this work. This will allow them to be confident that their result comes from a region or neighborhood of similar rather than dissimilar interpretation.

Although we contend that CLIFF can be applied to any type of trace evidence, in future work we hope to acquire more data sets to test CLIFF on. Also, direct comparison with other evaluation models will be investigated.

Bibliography

[1] CGG Aitken and D Lucy. Evaluation of trace evidence in the form of multivariate data. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS*, 53(Part 1):109–122, 2004.

[2] CGG Aitken, D Lucy, G Zadora, and JM Curran. Evaluation of transfer evidence for three-level multivariate data with the use of graphical models. *COMPUTATIONAL STATISTICS & DATA ANALYSIS*, 50(10):2571–2588, JUN 20 2006.

[3] JC Bezdek and LI Kuncheva. Some notes on twenty one (21) nearest prototype classifiers. In Ferri, FJ and Inesta, JM and Amin, A and Pudil, P, editor, *ADVANCES IN PATTERN RECOGNITION*, volume 1876 of *LECTURE NOTES IN COMPUTER SCIENCE*, pages 1–16. 2000.

[4] JC Bezdek and LI Kuncheva. Nearest prototype classifier designs: An experimental study. *INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS*, 16(12):1445–1473, DEC 2001.

[5] C.M Bishop. *Pattern Recognition and Machine Learning*. New York, NY, Springer, 2006.

[6] Jos Ramn Cano, Francisco Herrera, and Manuel Lozano. Stratification for scaling up evolutionary prototype selection. *Pattern Recognition Letters*, 26(7):953 – 963, 2005.

[7] Chin-Liang Chang. Finding prototypes for nearest neighbor classifiers. *Computers, IEEE Transactions on*, C-23(11):1179–1184, Nov. 1974.

[8] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, Jan 1967.

[9] B.V. Dasarathy. Minimal consistent set (mcs) identification for optimal nearest neighbor decision systems design. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(3):511–517, Mar 1994.

[10] V. Susheela Devi and M. Narasimha Murty. An incremental prototype set building technique. *Pattern Recognition*, 35(2):505 – 513, 2002.

[11] Richard O. Duda and Peter E.Hart. *Pattern classification and scene analysis*. A Wiley-Interscience Publication, New York: Wiley, 1973.

[12] Ian Evett. A quantitative theory for interpreting transfer evidence in criminal cases. *Applied Statistics*, 33(1):25–32, 1984.

[13] Ian Evett and John Buckleton. The interpretation of glass evidence. a practical approach. *Journal of the Forensic Science Society*, 30(4):215–223, 1990.

[14] Ian Evett and J. Lambert. Further observations on glass evidence interpretation. *Science and Justice*, 35(4):283–289, 1995.

[15] Christos Faloutsos and King-Ip Lin. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 163–174, New York, NY, USA, 1995. ACM.

[16] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[17] Utpal Garain. Prototype reduction using an artificial immune model. *Pattern Anal. Appl.*, 11(3-4):353–363, 2008.

[18] Salvador Garca, Jos Ramn Cano, and Francisco Herrera. A memetic algorithm for evolutionary prototype selection: A scaling up approach. *Pattern Recognition*, 41(8):2693 – 2709, 2008.

[19] D.M. Grove. Interpretation of forensic evidence using a likelihood ratio. *Biometrika*, 67(1):243–246, April 1980.

[20] P. Hart. The condensed nearest neighbor rule (corresp.). *Information Theory, IEEE Transactions on*, 14(3):515 – 516, may 1968.

[21] O. Jalali, T. Menzies, and M. Feather. Optimizing requirements decisions with keys. In *Proceedings of the PROMISE 2008 Workshop (ICSE)*, 2008. Available from `http://menzies.us/pdf/08keys.pdf`.

[22] I. Jolliffe. *Principal component analysis*. Springer-Verlag, 175 Fifth Avenue, NY, USA, 2002.

[23] N. Karslake, S. Lewis, and W. Bronswijk. Characterisation of automotive paint clear coats by atr-fr-ir with subsequent chemometric analysis. 2009.

[24] SW Kim and BJ Oommen. A brief taxonomy and ranking of creative prototype reduction schemes. *PATTERN ANALYSIS AND APPLICATIONS*, 6(3):232–244, DEC 2003.

[25] T. Kohonen. Improved versions of learning vector quantization. pages 545 –550 vol.1, jun 1990.

[26] Teuvo Kohonen and Panu Somervuo. Self-organizing maps of symbol strings. *Neurocomputing*, 21(1-3):19 – 30, 1998.

[27] RD Koons and J Buscaglia. Interpretation of glass composition measurements: the effects of match criteria on discrimination capability. *JOURNAL OF FORENSIC SCIENCES*, 47(3):505–512, MAY 2002.

[28] RD Koons and JA Buscaglia. The forensic significance of glass composition and refractive index measurements. *JOURNAL OF FORENSIC SCIENCES*, 44(3):496–503, MAY 1999.

[29] F Korn, BU Pagel, and C Faloutsos. On the "dimensionality curse" and the "self-similarity blessing". *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 13(1):96–111, JAN-FEB 2001.

[30] Y Li, M Xie, and T Goh. A study of project selection and feature weighting for analogy based software cost estimation. *Journal of Systems and Software*, 82:241–252, 2009.

[31] DV LINDLEY. PROBLEM IN FORENSIC-SCIENCE. *BIOMETRIKA*, 64(2):207–213, 1977.

[32] Committee on Identifying the Needs of the Forensic Sciences Community;Committee on Applied and National Research Council. Theoretical Statistics. *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055, 2009.

[33] Allan Seheult. On a problem in forensic science. *Biometrika*, 65(3):646–648, December 1978.

[34] CJ Veenman and MJT Reinders. The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 27(9):1417–1429, SEP 2005.

[35] K. Walsh, J. Buckleton, and C. Triggs. A practical example of the interpretation of glass evidence. *Science and Justice*, 36(4):213–218, 1996.

[36] G. Zadora. Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian Network approaches. *ANALYTICA CHIMICA ACTA*, 642(1-2, Sp. Iss. SI):279–290, MAY 29 2009.

[37] G. Zadora and T. Neocleous. Likelihood ratio model for classification of forensic evidence. *ANALYTICA CHIMICA ACTA*, 642(1-2, Sp. Iss. SI):266–278, MAY 29 2009.