Strangeness Minimisation Feature Selection with Confidence Machines

Tony Bellotti, Zhiyuan Luo, and Alex Gammerman

Computer Learning Research Centre, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
zhiyuan@cs.rhul.ac.uk

Abstract. In this paper, we focus on the problem of feature selection with confidence machines (CM). CM allows us to make predictions within predefined confidence levels, thus providing a controlled and calibrated classification environment. We present a new feature selection method, namely Strangeness Minimisation Feature Selection, designed for CM. We apply this feature selection method to the problem of microarray classification and demonstrate its effectiveness.

1 Introduction

Confidence machine (CM) is a framework for constructing learning algorithms that predict with confidence. In particular, we use CM to produce hedged predictions with accuracy controlled by predefined confidence level. Predictions using CM are not only accurate but also, unlike many conventional algorithms, well-calibrated [2,5]. This method is ideally suited to the problem of providing measured and controlled risk of error for microarray analysis. In this paper, we focus on the problem of feature selection with CM and describe a new feature selection method, namely Strangeness Minimisation Feature Selection, designed for CM. We apply this feature selection method to the problem of microarray classification and demonstrate its effectiveness.

2 Confidence Machines

CM is a general learning framework for making well-calibrated predictions, in the sense that test accuracy is controlled. Intuitively, we predict that a new example (eg microarray data for a new patient with unknown diagnosis) will have a label (eg disease diagnosis) that makes it similar to previous examples in some specific way, and we use the degree to which the specified type of similarity holds within the previous examples to estimate confidence in the prediction. Formally, CMs work by deriving a p-value based on a *strangeness measure*. A strangeness measure is a function that provides an indication of how strange or typical a new example is in relation to a given sequence of examples.

Once a strangeness measure A is defined, we can compute the p-value for a new example. Given a sequence of labelled examples for training $(z_1, ..., z_{n-1})$

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 978–985, 2006.

where each example z_i consists of an object $\mathbf{x}_i$ and its label y_i, and a new example $\mathbf{x}_n$ with label withheld, we calculate p-values for each $y \in Y$, where Y is the set of all possible class labels, as

$$p_y = \frac{|\{i : i \in \{1, ..., n\}, \alpha_i \geq \alpha_n\}|}{n} \tag{1}$$

where $\alpha_i = A(z_i, (z_1, ..., z_n))$ and $z_n = (\mathbf{x}_n, y)$ for $i \in \{1, ..., n\}$. So, p_y is computed from the sequence $(z_1, ..., z_{n-1}, (\mathbf{x}_n, y))$. Clearly $\frac{1}{n} \leq p_y \leq 1$. The p-value p_y gives a measure of typicalness for classifying $\mathbf{x}_n$ with label y. The use of p-values in this context is related to the Martin-Löf test for randomness [3]. Indeed, the p-values generated by CM form a valid randomness test under the iid assumption.

A region prediction for $\mathbf{x}_n$ is then computed as $R = \{\{y : p_y > \delta\} \cup \arg\max_y(p_y)\}$ where $1 - \delta$ is a confidence level supplied prior to using CM (and $\arg\max$ here returns the *set* of arguments giving the maximum value). This region prediction will always predict the label with the highest p-value, but may also hedge this prediction by including any other label with a p-value greater than δ. If $|R| = 1$, i.e. only one label is predicted, it is called a *certain prediction*. Otherwise, if $|R| > 1$, it is an *uncertain prediction*. Clearly, certain predictions are more efficient than uncertain ones, so an objective of learning with region predictions is to make as many certain predictions as possible. A region prediction is correct if the true label for the example is a member of the predicted region. Otherwise it is an error. Region predictions are well-calibrated, in the sense that we expect the number of errors for k predictions to be approximately less than or equal to $k\delta$ [11].

CM can operate as an on-line or off-line learner. In the on-line learning setting the examples are presented one by one. Each time, the classifier takes an object to predict its label. Then the classifier receives the true label from an ideal teacher and goes on to the next example. The classifier starts by observing $\mathbf{x}_1$ and predicting its label $\hat{y_1}$. Then the classifier receives feedback of the true label y_1 and observes the second object $\mathbf{x}_2$, to predict its label $\hat{y_2}$. The new example $(\mathbf{x}_2, y_2)$ is then included in the training set for the next trial. And so on. At the nth trial, the classifier has observed the previous examples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n-1}, y_{n-1})$ and the new object $\mathbf{x}_n$ and will predict $\hat{y_n}$. We expect the quality of the predictions made by the classifier to improve as more and more examples are accumulated. In the off-line learning setting, the classifier is given a training set $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$ which is then used to make predictions on new unlabelled examples $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, ..., \mathbf{x}_{n+k}$. For on-line learning, it is possible to prove that CM is well-calibrated, in the sense that the test errors form a Bernoulli sequence with parameter δ if we assume identically and independently distributed (iid) data. Calibration to confidence level is a useful property of CM allowing direct control of risk of error. For example, CM has been applied successfully to the problem of reliable diagnosis from microarray data and proteomics patterns [2,5]. In order to achieve calibration, the strangeness measure needs to be *exchangeable* in the sense that the strangeness value α_i is not dependent on the order of the sequence of labelled examples given to Equation (1) [11].

3 Feature Selection

Most microarray gene expression data sets suffer from the dual problem of low sample size and high-dimensional feature space [8]. It is well known that dimension reduction is necessary when high dimensional data is analysed. This is because,

1. large number of features may cause over-fitting, if they are not relevant features, and if the underlying distributions are not estimated accurately,
2. large number of features makes it difficult to design a classifier due to time and space complexity issues.

Previous work with gene expression data has shown that the use of feature selection during classification can lead to improvements in performance, in light of these problems; eg Yeoh et al [12].

Feature selection is the process of selecting a subset of features from a given space of features with the intention of meeting one or more of the following goals:

- choose the feature subset that maximises the performance of the learning algorithm;
- minimise the size of the feature subset without reducing the performance of a learning algorithm on a learning problem significantly;
- reduce the requirement for storage and computational time to classify data.

There are three general methods for feature selection: filters, wrappers and embedded feature selection. The filter method employs a feature ranking function to choose the best features. For example, the signal to noise ratio (SNR) is a ranking function that scores a feature by how well it is a signal for the classification label. Wrapper methods are general purpose algorithms that search the space of feature subsets, testing performance of each subset using a learning algorithm. Some learning algorithms include an embedded feature selection method. Selecting features is then an implicit part of the learning process. This is the case, for example, with decision learners like ID3.

It is possible to derive feature selection from within the CM framework. In this paper, we propose a new method called Strangeness Minimisation Feature Selection (SMFS) that selects the subset of features that minimise the overall strangeness values of a sequence of examples. The intuition for this approach is that reducing overall strangeness implies an increase in conformity amongst the examples in the sequence. Therefore, the set of features that minimise overall strangeness are most relevant to maximising conformity between training examples.

3.1 Strangeness Minimisation Feature Selection

The SMFS goal is defined in relation to any strangeness measure. Let $\tilde{A}$ be an exchangeable measure. The optimal strangeness minimisation feature subset S_0 of size t for a sequence $(z_1, ..., z_n) \in Z^n$ with feature space F is the SMFS goal,

$$S_0 = \arg\min_{S \in G} \sum_{i=1}^{n} \tilde{A}(z_i, (z_1, ..., z_n), S) \tag{2}$$

where $G = \{R : R \subseteq F, |R| = t\}$. We establish that this is an exchangeable feature selection function and can be implemented in CM. In common with wrapper feature selection methods, it requires a search over the space of all subsets of F, although restricted to subsets of size t. For m features, in the typical case when $t < m$, this search has computational complexity of $O(m^t)$. The number of selected features t does not need to be very large for this to become intractable. Fortunately, practical implementations of SMFS are possible if we restrict our attention to a subclass of linear strangeness measures. The problem becomes tractable without the need for ad-hoc heuristics that are often required with wrapper methods. Within CM framework, we can still implement useful versions of CM based on learning algorithms such as nearest centroid and SVM. We find that SMFS is a principled, broad and practical feature selection framework, for which distinct feature selection methods are determined by strangeness measures.

The function $\tilde{A} : \mathbf{Z} \times \mathbf{Z}^n \times 2^F \to \mathbf{R}$ is a linear measure based on the transformation function $\phi : \mathbf{Z} \times F \times \mathbf{Z}^n \to \mathbf{R}$ if

$$\tilde{A}(z_i, (z_1, ..., z_n), S) = \sum_{j \in S} \phi(z_i, j, (z_1, ..., z_n)) \tag{3}$$

for all $z_i \in Z$, and $S \subseteq F$.

In CM, strangeness examples are computed as α-strangeness values for each example, see Eq (1). By using linear strangeness measures, we can compute strangeness values for each feature. We call them β-strangeness measures. The β-strangeness value for feature j is defined as

$$\beta_j = \sum_{i=1}^{n} \phi(z_i, j, (z_1, ..., z_n)). \tag{4}$$

We reformulate the SMFS goal to minimise the sum of β-strangness measure values across subsets of features of size t,

$$S_0 = \arg\min_{S \in G} \sum_{j \in S} \beta_j \tag{5}$$

where $G = \{R : R \subseteq F, |R| = t\}$. It is easy to see that this minimum is computed from the set of features giving the smallest β-strangeness values. Hence, the SMFS goal can be solved by

1. computing β-strangeness values for each feature,
2. sorting the β-strangeness values in ascending order and choosing the top t.

Clearly, this solution is tractable and is a great improvement on the computational complexity for SMFS in general.

We now consider implementing SMFS using two examples of linear strangeness measures and construct strangeness measures based on the Nearest Centroid (NC) classifier [8] and linear classifier. In both cases, we derive an exchangeable transformation function and the corresponding β-strangeness measure.

For the NC classifier, a linear strangeness measure can be given using specifically one of family of Minkowski distance metrics,

$$\tilde{A}_{NC}((x_0, y_0), (z_1, ..., z_n), S) = \sum_{j \in S} (\frac{|x_{0j} - \underline{\mu}_{(y_0)j}|}{\sigma_j})^k \tag{6}$$

where $\underline{\mu}_{(y_0)j}$ is the within class mean for label y_0 and σ_j variance for feature j based on the sequence $(z_1, ..., z_n)$. If we consider NC with the usual Euclidean distance measure ($k = 2$), the β-strangeness measure corresponding to the NC linear strangeness measure $\tilde{A}_{NC}$ is

$$\beta_j = \frac{\sum_{y \in Y} (|C_y| \sigma^2_{(y)j})}{\sigma^2_j} \tag{7}$$

where $|C_y|$ is the number of examples with label y.

We have described a strangeness measure for SVM based on the Lagrange multipliers defined in the dual form optimisation problem for SVM [5]. Although these work for SVM, they cannot be applied to other linear classifiers in general. Also, the range of strangeness values that are output is small since all nonsupport vector examples will have a strangeness value of zero. This means strangeness cannot be measured between these examples. We define a new strangeness measure which resolves both these difficulties, by measuring the distance between an example and the separating hyperplane. The linear classifier strangeness measure is the function $\tilde{A}_{LC}$ defined for all (x_0, y_0) and $(z_1, ..., z_n)$ as

$$\tilde{A}_{LC}((x_0, y_0), (z_1, ..., z_n), S) = -y_0(\sum_{j \in S} w_j x_{0j} + b) \tag{8}$$

where the hyperplane (w, b) has been computed using a linear inductive learner based on the sequence $(z_1, ..., z_n)$. We can take the threshold b as an extra weight on a new constant-valued feature without loss of generality. The β-strangeness measure corresponding to the above linear classifier strangeness measure $\tilde{A}_{LC}$ is

$$\beta_j = -w_j \sum_{i=1}^{n} y_i x_{ij}. \tag{9}$$

4 Experiments

Our experiments are based on two public databases of Affymetrix HG-U133A oligonucleotide microarrays for children with acute leukaemia. These microarrays provide gene expression measurements on over 22,000 gene probes for human

genes. The data are given with an established clinical subtype classification for each microarray. Each example is classified as either ALL or AML and then as a specific subtype within each of these groups. Therefore each patient is represented as a vector of expression levels and a subtype label.

Dataset A comprises 84 microarrays sourced from Royal London Hospital, UK [10]. There are 62 ALL and 22 AML examples and the data is available at `nostradamus.cs.rhul.ac.uk/MicroArray`. Dataset A was normalised so that each example had the same mean expression level (across all gene probes) and a log transform applied prior to classification. Dataset B is a combination of 132 ALL cases [7] and 130 AML cases [6] which are both sourced from St. Jude Children's Research Hospital, USA and are available at `www.stjuderesearch.org/data`.

Strangeness Minimisation Feature Selection (SMFS) is a new feature selection method proposed in this paper. The effectiveness of these two new linear strangeness measures is tested on the two microarray datasets and compared with other feature selection techniques, such as using the SNR filter. SVM is used as the linear classifier without kernels. Both test accuracy and and efficiency which is the ratio of certain predictions are measured. For predictions to be useful, the large majority of region predictions need to be certain.

Table 1. Dataset A: On-line CM with different feature selection methods

Feature selection method	Accuracy	Efficiency
None	1	0.018
SNR	0.976	0.607
ANOVA	0.988	0.702
SAM	1	0.750
SMFS	1	0.738

The NC strangeness measure given in Equation (7) for SMFS was applied to Dataset A classifying for subtypes ALL or AML. The online setting was used to determine the effectiveness of classification over the time. CM was run without feature selection and then with SMFS taking the top $t = 40$ features. Both were run in the online setting with a confidence level of 95%. SMFS was also contrasted with several established feature selection techniques for microarray analysis: the SNR filter, analysis of variance (ANOVA) as available in the GeneSpring 5.0.3 and the more sophisticated Significance Analysis of Microarrays (SAM) for reducing the false discovery rate [9]. The top $t = 40$ genes were chosen for each method and a 95% confidence level was set. Table 1 shows results at the last trial. SMFS gives clearly better results than SNR, in terms of less errors and uncertain predictions, is slightly better than ANOVA and is comparable with SAM. Both ANOVA and SAM were run on the entire data set prior to classification so we might expect some selection bias with these experiments [1].

CM was run with the linear classifier strangeness measure define in Equation (9), based on SVM. The same algorithm was applied to Dataset B using 10 cross validation. Since the linear classifier SVM is binary classifier, the binary

Table 2. Dataset B: Off-line CM results

	Confidence level					
	97.5%		95%		90%	
Method	Accuracy	Efficiency	Accuracy	Efficiency	Accuracy	Efficiency
SVM						
No FS	0.989	0.034	0.966	0.065	0.939	0.160
SNR	0.989	0.496	0.977	0.710	0.973	0.966
SMFS	0.989	0.996	0.989	1	0.989	1
NC						
SMFS	0.992	0.992	0.992	0.992	0.992	0.992

classification task of discriminating between subtypes ALL and AML was undertaken. It is contrasted with applying CM without feature selection and with using the SNR filter feature selection, instead of SMFS. Results are shown in Table 2 for different confidence levels. This shows that using SMFS yields the best performance in terms of accuracy and efficiency and CM is well-calibrated in all cases. The results for CM based on the NC strangeness measure are also shown for 10 cross validation. Both NC and SVM perform well with NC giving slightly higher accuracy but less efficiency. Standardisation was applied as a preprocessing step prior to classification using the SVM linear classifier strangeness measure [4].

5 Conclusions

Feature selection is important for successful microarray classification. We have described the recently developed feature selection method, namely strangeness minimisation feature selection, designed for confidence machines. Our experiments demonstrate that SMFS works well.

Acknowledgments

The authors would like to thank Frederik van Delft and Vaskar Saha of Cancer Research UK Children's Cancer Group for their discussions. The Confidence Machine project has received financial support from the following bodies: EPSRC through studentship; MRC through grant S505/65.

References

1. Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. In *Proceedings of the National Academy of Sciences of the USA*, volume 99, pages 6562–6566, May 2002.
2. T. Bellotti, Z. Luo, A. Gammerman, F.W. van Delft, and V. Saha. Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. *International Journal of Neural Systems*, 15(4):247–258, 2005.

3. Alex Gammerman and Volodya Vovk. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 287:209–217, 2002.
4. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
5. Zhiyuan Luo, Tony Bellotti, and Alex Gammerman. Qualified predictions for proteomics pattern diagnostics with confidence machines. In *Intelligent Data Engineering and Automated Learning - IDEAL 2004 (LNCS 3177)*, pages 46–51. Springer, 2004.
6. Mary E. Ross, Rami Mahfouz, Mihaela Onciu, Hsi-Che Liu, Xiaodong Zhou, Guangchun Song, Sheila A. Shurtleff, Stanley Pounds, Cheng Cheng, Jing Ma, Raul C. Ribeiro, Jeffrey E. Rubnitz, Kevin Girtman, W. Kent Williams, Susana C. Raimondi, Der-Cherng Liang, Lee-Yung Shih, Ching-Hon Pui, and James R. Downing. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, 104(12):3679–3687, 2004.
7. Mary E. Ross, Xiaodong Zhou, Guangchun Song, Sheila A. Shurtleff, Kevin Girtman, W. Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C. Raimondi, Noel Lenny, Anami Patel, and James R. Downing. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102:2951–2959, 2003.
8. Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Association of Science (USA)*, 99(10):6567–6572, May 2002.
9. Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significant analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences of the USA*, volume 98, pages 5116–5121, 2001.
10. F.W. van Delft, T. Bellotti, M. Hubank, T. Chaplin, N. Patel, M. Adamaki, D. Fletcher, I. Hann, L. Jones, A. Gammerman, and V. Saha. Hierarchical clustering analysis of gene expression profiles accurately identifies subtypes of acute childhood leukaemia. *British Journal of Cancer*, 88:54–54, 2003.
11. Vladimir Vovk. On-line confidence machines are well-calibrated. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 187–196. IEEE Computing, 2002.
12. Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouza, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pedriatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, March 2002.