Strangeness-based Feature Weighting and Classification of Gene Expression Profiles

Haifeng Shao
Genetics
Case Western Reserve Univ.
Cleveland, Ohio 44120, USA
hxs61@cwru.edu

Bei Yu
Kellogg School of Management
Northwestern University
Evanston, IL 60208, USA
bei.yu@gmail.com

Joseph Nadeau
Genetics
Case Western Reserve Univ.
Cleveland, Ohio 44120, USA
jhn4@cwru.edu

ABSTRACT

Achieving high classification accuracy is a major challenge in the diagnosis of cancer types based on gene expression profiles. These profiles are notoriously noisy in that a large number of genes might be irrelevant to or weakly associated with disease phenotypes such as tumors. Assigning different weights to genes could decrease or diminish the influences of those "noisy" signals, and thereby improve classification accuracy. We propose an intuitive and simple approach to cancer classification with feature weighting. Our strangeness-based feature weighting method learns weights for different genes based on their classification performance. Those genes with large weights can be used as discriminative genes. We demonstrate that our implementation of k-NN classifier achieved high classification accuracy on two benchmark cancer data sets. In the case of relatively low accuracy, the proposed method could be used as a feature filter. With combined feature weighting and AdaBoost, we achieved a better classification accuracy (100%) than using strangeness-based k-NN alone.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

Keywords

Strangeness, Feature weighting, Gene expression, Cancer classification

1. INTRODUCTION

Accurate cancer diagnosis is important and challenging. It is essential to make correct diagnosis in order to provide proper medical treatments. Traditional techniques, such as examination of histological features, cytogenetic and morphological appearances, and immunohistochemistry, either are subjective or do not always give a definitive diagnosis [2].

SAC'08 March 16-20, 2008, Fortaleza, Ceara, Brazil

Gene expression profiling has been a great resource for cancer diagnosis. Differentially expressed genes from different tissue and tumor types provide predictors or signatures of certain cancer classes or subclasses, which in turn can be used as the basis of classification. As such, gene expression profiles are more objective than subjective. Building accurate classifiers and finding relevant genes are challenging in that in practice we have thousands of genes for a much smaller number of noisy samples, which is usually in the order of tens.

Statistical learning methods, such as nearest shrunken centroids (NSC) [4], principle component analysis (PCA) and supervised learning algorithms such as support vector machines (SVM) and Neural Networks (NN) [1, 2, 3] have been applied for these purposes. While a PCA-based classifier does not provide specific information on individual genes, most methods often suggest a list of genes which might be involved in certain pathways related to the cancer stage. It is reasonable to assume that different genes have different impacts on different tissues, which implies that different genes have different importance in terms of discriminating phenotypes such as tumors. Instead of treating selected genes equally, gene weighting might be a more appropriate approach for finding gene signatures. Furthermore, parametric classifiers such as SVMs and NNs usually demand estimation of many parameters and make certain assumptions of expression data that are hard to verify in small data sets. On the other hand, k-NN classifiers as nonparametric classifiers, often have superior classification ability for high-dimensional data, such as in text mining [7]. However, k-NN does not associate specific weights to features.

Here we propose a method – strangeness-based feature weighting algorithm (for which feature selection can be considered as a special case). Strangeness of a sample is the ratio of the sum of k nearest distances from the same class to that from other classes. It measures the relative relationship of the sample to those from the same class and to those from other classes. The average strangeness of a sample set is low if each sample is close to samples in the same class and far from samples in other classes. Our method optimizes a feature weight vector such that the average strangeness of the whole sample set is minimal. This method makes no assumptions about the distribution of the data and is easy to implement. Our implementation with k-nearest neighbor (k-NN) classifier showed that this method provided comparable or improved accuracy to those based on more com-

plicated methods. We also used the proposed method as a preprocessing step to select top weighted features, and then implemented AdaBoost to build a collection of weak classifiers. The combined final classifier further improved the classification accuracy.

2. METHODS

2.1 Strangeness measure

Several strangeness measurements have been proposed under the context of machine learning [9, 10, 11, 15]. Here we use k-nearest neighbor (k-NN) strangeness, as described in [9], because a non-parametric method is more suitable in practice and we do not wish to make any assumptions about the sample distribution. Consider N tissue samples belonging to different classes. For each sample, we have P transcripts (genes) as features. Let x_{ij}^c be the expression level of jth gene from ith sample that belongs to class c. Let d_{il}^c be the Euclidean distance between samples i and l where i and l belong to the same class c. Similarly let d_{il}^{-c} be the Euclidean distance between samples i and l where i belongs to class c and l belong to any other class but c. We can sort those ds from samples belonging to the same class as sample i and those from different classes. For each sample i, we define its strangeness with respect to putative class c as:

$$\alpha_i = \frac{\sum_{l=1}^{k} d_{il}^c}{\sum_{l=1}^{k} d_{il}^{-c}} \tag{1}$$

2.2 Naive strangeness-based feature weighting

Even though strangeness measures the relative similarity of samples from the same class to that from other classes, it treats each feature dimension in an unbiased manner. In cancer diagnosis based on gene profiles, genes have different associations with different cancer types. Some might not have an association with any cancer type, while others might have a strong association with particular cancer type(s). Therefore, different genes should have different influences on the distance measurement of samples. To incorporate this concept, we introduce a weight vector $\omega \in R^P$ over the P variables (features), $\omega_p \geq 0, p = 1, \cdots, P$. The strangeness value of example $\mathbf{x}_i$ is represented as a function of the weight vector $\boldsymbol{\omega}$:

$$\alpha_i^{\omega} = \frac{\sum_{l=1}^{k} d(\omega)_{il}^c}{\sum_{l=1}^{k} d(\omega)_{il}^{-c}} \tag{2}$$

where

$$d(\omega)_{il} = \|\mathbf{x}_i - \mathbf{x}_l\|_{\omega} = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{lp})^2 \omega_p^2} \tag{3}$$

Given a training set $\mathbf{x}_1, \cdots, \mathbf{x}_N$ and the weight vector ω, the evaluation function can be defined as

$$\Phi(\omega) = \frac{1}{N} \sum_{\mathbf{x}_i} \alpha_i^{\omega} \tag{4}$$

The evaluation function represents the average separability of different classes. The smaller value the objective function is, the more separable the samples are. To achieve high classification accuracy, our objective is to find a set of weights, that is weight vector ω, such that it minimizes the evaluation function $\Phi(\omega)$ as defined in equation 4, therefore achieves high classification accuracy. The selected feature set can be defined by using a threshold on ω which is similar to the *RELIEF* algorithm [13]. Since $\Phi(\omega)$ is smooth almost everywhere, a gradient descent method is used to minimize it. With some algebra, the gradient of $\Phi(\omega)$ with respect to weight ω_i is:

$$(\nabla\Phi(\omega))_i = \frac{1}{N} \sum_{\mathbf{x}_j} \left(\frac{\zeta_1(\omega)\varphi_2(\omega) - \zeta_2(\omega)\varphi_1(\omega)}{\varphi_2^2(\omega)}\right)\omega_i \tag{5}$$

where $\varphi_1(\omega) = \sum_{l=1}^{k} d(\omega)_{jl}^c$, $\varphi_2(\omega) = \sum_{l=1}^{k} d(\omega)_{jl}^{-c}$, $\zeta_1(\omega) = \sum_{l=1}^{k} \frac{(x_{ji} - x_{li}^c)^2}{d(\omega)_{jl}^c}$ and $\zeta_2(\omega) = \sum_{l=1}^{k} \frac{(x_{ji} - x_{li}^{-c})^2}{d(\omega)_{jl}^{-c}}$.

Here, we use a stochastic gradient descent over the evaluation function $\Phi(\omega)$, and develop an iterative strangeness-based feature weighting algorithm as in Algorithm 2.2, where ρ is the learning rate and T is the number of iterations. For N training data points in P dimensional space, the computational complexity of Algorithm 2.2 is $O(PTN)$. By using kd trees to extract k nearest neighbors, we can speed up the algorithm leading to a reduced computational complexity of only $O(NTlogP)$.

Iterative Strangeness-based Feature Weighting with Stochastic Search

1. Initialize $\omega = (1, \cdots, 1)^P$.
2. For $t = 1, \cdots, T$
   - Pick randomly any data point $\mathbf{x}_t$ from training data set.
   - Calculate the strangeness of $\mathbf{x}_t$ with respect to ω.
   - For $i = 1, \cdots, P$, calculate
   $$\delta_i = \left(\frac{\zeta_1(\omega)\varphi_2(\omega) - \zeta_2(\omega)\varphi_1(\omega)}{\varphi_2^2(\omega)}\right)\omega_i$$
   - Update $\omega(t) = \omega(t) - \rho\Delta$
   - Check error rate, if it increases, stop.
3. For the feature selection purpose, the selected feature set is $\{i|\omega_i > \tau\}$, where τ is a threshold determined through validation.

2.3 K-NN classifier with strangeness-based feature weighting

The incorporation of strangeness-based feature weighting and k-NN classifier is straightforward since our definition of strangeness is k-NN based. At the testing stage, for each test sample we compute the strangeness value of this sample with putative labels. To predict the class to which the test sample belongs, we could simply predict the label associated with the smallest strangeness value.

2.4 Boosting with strangeness-based feature selection (FS)

In multiclass classification problems, especially when the classes are not well separated, a simple individual classifier

Table 1: Test result comparison of Leukemia data ($k = 3$, all features were used for the first 3 methods)

Method/Class	ALL	AML	Accuracy
	predicted (true)	predicted (true)	
k-NN	20(20)	10(14)	88%
SVM	20(20)	11(14)	91%
RF	20(20)	6(14)	77%
k-NN+FS	20(20)	12(14)	94%
k-NN+FS+Boosting	20(20)	14(14)	100%

such as k-NN classifier can not provide superb results. For such cases, we develop a boosting frame with strangeness-based feature selection as a preprocessing step. Our learning model is based on the AdaBoost algorithm [8]. AdaBoost includes the possibility of assigning different weights w_i to positive and negative training samples, and delivers a classifier $h(x) : x \rightarrow y$ to predict whether a given sample has the class label c. Detailed information about AdaBoost can be found in [8]. For multiclass problems, Freund and Schapire proposed two algorithms: AdaBoost.M1 and AdaBoost.M2. AdaBoost.M2 algorithm was used in this work (detailed algorithm will be reported elsewhere).

3. RESULTS

The strangeness-based feature weighting algorithm with k-NN classifier was applied to two microarray data sets. The results are presented and compared with other classifiers such as SVM and random forests (RF). We also show the results after boosting for comparison purposes.

3.1 Classification of acute leukaemia samples

The first data set analyzed is from leukemia samples [1]. This data set has become a benchmark for classification and clustering algorithms [3, 4, 5]. The data set was divided into two subsets, a training set and a test set. There are 38 samples in the training set with 11 acute myeloid leukemia (AML) and 27 acute lymphoblastic leukemia (ALL). Among those 27 ALL samples, there are 19 B-cell ALLs and 8 T-cell ALLs. The test set consists 34 samples with 20 ALLs and 14 AMLs. The experiment measured 7129 genes using Affymetrix oligonucleotide microarrays. The original analysis only considered a 2-class problem [1]. For comparison, we implemented five classifiers on this data set: k-NN (k=3), SVM with linear kernel, RF, k-NN with strangeness-based feature selection (FS), and k-NN strangeness-based FS with boosting. k-NN classifier was implemented in R using {class} package. SVM was implemented in R with {e1071} package. Linear kernel was used with leave-one-out cross validation. The trained SVM classifier consisted of 31 support vectors. The overall validation accuracy was 91%. RF was implemented in R with {randomForest} package and used 500 trees. The prediction results on the test samples are shown in Table 1. We show the best result with the top 20 genes selected for the last two models.

The k-NN classifier with strangeness-based feature selection achieved better prediction accuracy than k-NN classifier without feature selection. In general (100 runs), the proposed method misclassified $2 \sim 4$ samples. This indicates that with many fewer features (genes), we maintained the same predictive power using strangeness-based feature selection. This is important in high dimensional data analysis because it saves storage space and in turn speeds up the processing time without sacrificing the ultimate outcome. We used 3-fold cross-validation during training to avoid overfitting. Most of our top 20 genes were identified as relevant to leukemia biology in other studies [1, 5], such as MPO (M19507) and IL-8 precursor (Y00787). Our algorithm also identified several genes that were also strong predictors of ALL-B-Cell and ALL-T-Cell, such as X03934 which is the T-cell antigen receptor gene and U05259 which is the B-cell antigen receptor complex-associated protein alpha-chain precursor (Ig-alpha) (Figure 1).

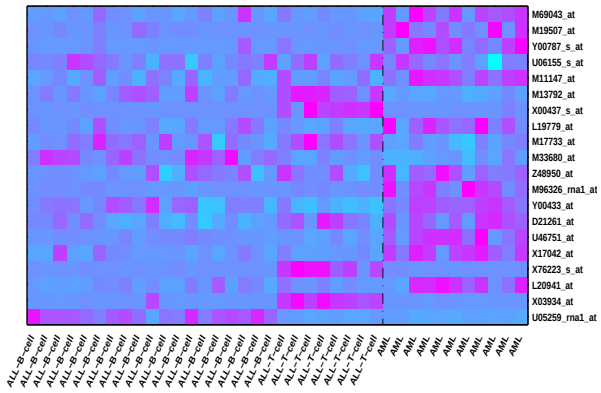

Figure 1: Heatmap of the top 20 genes selected by strangeness-based feature weighting method from the leukemia training dataset. Gene M19507 was assigned the largest weight. Each column represents one sample with its name at the bottom.

The strangeness values for the training and test sets are shown in Figures 2 and 3 respectively. For each sample, we computed the strangeness values with putative classes. The predicted class is the one associated with the smallest strangeness value. For instance, in Figure 2, the first sample is ALL (true label) and was predicted to be ALL because this sample had a smaller strangeness value if it belongs to ALL class than that of AML class. As shown in Figure 3, four test samples were misclassified by using strangeness-based k-NN classifier. To improve the classification accuracy, we implemented the boosting framework as described in the previous section. The predicted class label is the one with the largest output from the final classifier. On average (100 runs), with strangeness and boosting, the prediction accuracy was improved from 88% to 97%. The best classification accuracy of the final classifier was 100% (not shown).

3.2 Diagnosis of small round blue cell tumors of childhood

In this data set, the Small Round Blue Cell Tumors (SRBCT) of childhood were classified into four classes, namely Burkitt lymphoma (BL, a subset of non-Hodgkin lymphoma), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). 88 samples were divided into a training ($N = 63$) and a test set ($N = 25$). Each sample has 2308 genes after filtering for a minimal level of expression [2], which was used in this work. Among these

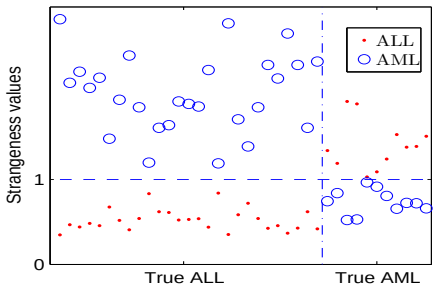

Figure 2: Leukemia training set.

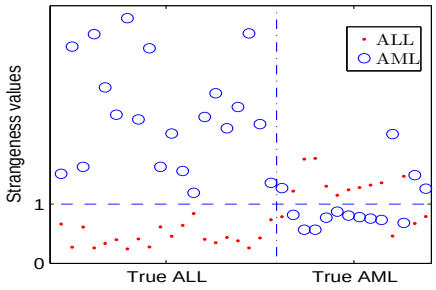

Figure 3: Leukemia test set.

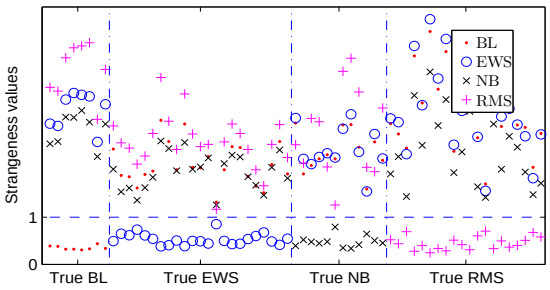

Figure 5: SRBCT training set.

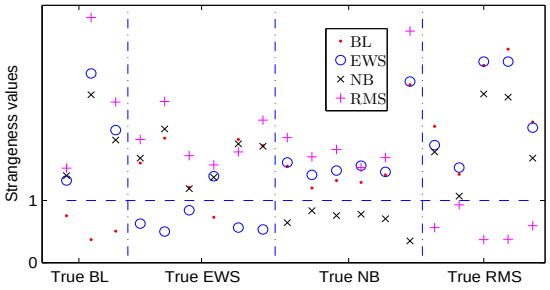

Figure 6: SRBCT test set.

25 test samples, five are non-SRBCTs. In many medical classification cases, such as here, we encounter samples that have novel class labels. In other words, some classes were never seen at the training stage. In such cases, it is important to make no-call in order to avoid wrong diagnosis and treatment. In this case, we incorporated measurements of credibility and confidence levels [9, 11, 12] for diagnostic purposes. To estimate p-value as described, we chose $f(\alpha) = \alpha$, which is a monotonic non-decreasing function with $f(0) = 0$, as required. Confidence was defined as 1 minus the second largest p-value. Prediction was made by the largest p-value.

The heatmap of top 20 genes with largest weights (training set) is shown in Figure 4. 100% of the training samples were correctly assigned to their labels (Figure 5). With the blinded test set, we only had one sample (Test# 20) which

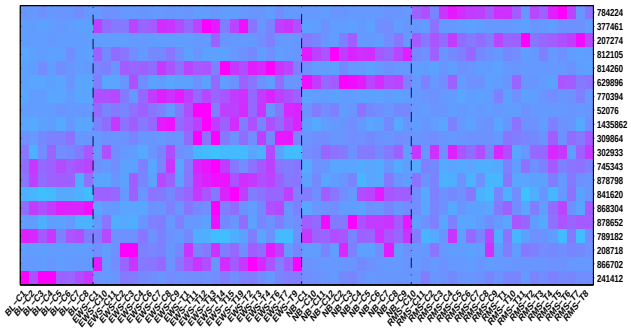

Figure 4: Heatmap of the top 20 genes selected by strangeness-based feature weighting method from the SRBCT training dataset.

would have been misclassified (Figure 6), but because of its low credibility, the diagnosis was no-call, meaning the classifier was not able to assign the test sample to any of the classes encountered at the training stage. We also correctly made no-call decisions for all of five non-SRBCT samples (not shown). Our methods yielded similar results to published methods [2, 4]. For comparison purposes, we implemented k-NN, SVM, and RF on 20 non-SRBCT samples with all features. The trained SVM consisted of 59 support vectors and its prediction accuracy from leave-one-out-cross-validation was 98.4%. RF used 1000 trees. 100% accuracy was obtained with k-NN, SVM, k-NN+FS and k-NN+FS+Boosting, and 95% with RF.

Our method also identified several genes that were differentially expressed in specific cancer types. For example, the largest weight was assigned to feature 784224 (Figure 4), which is fibroblast growth factor receptor 4 (FGFR4) and has a potential role in tumor growth and was preferentially expressed in RMS [2, 4]. Another gene, MIC2 (feature 1435862) was also among the top 10 weighted genes. This gene was highly expressed only in EWS [6]. These results demonstrated the effectiveness of the proposed method in terms of identifying potential gene (molecular) signatures to cancer types. When boosting was implemented, 100% classification accuracy was achieved (data not shown).

4. DISCUSSION

Improving tumor classification accuracy is a major challenge in cancer diagnosis and treatment. Recently, the use of gene expression profiles as features to classify different cancer types has attracted much attention because of its objective nature [1, 2, 5, 4, 3]. In this work, we described a new and simple feature weighting method based on a strangeness measure. We integrated it with the non-parametric k-NN

classifier for classification purposes. Only one parameter, k, needs to be tuned. The classification results from two benchmark data sets were comparable to or better than more complicated classifiers such as SVM, RF and NN. With boosting, the classification accuracy was improved to 100%. However, boosting introduced several parameters into the algorithm, such as τ in AdaBoost and the sets of features to select, which in turn added to the complexity of the algorithm and is therefore a tradeoff with improved accuracy.

The proposed method assigned weights to genes based on their discrimination ability. In the leukemia example, our proposed method assigned largest weights to genes such as MPO and IL-8 precursor, which were also identified in other studies [1]. In the SRBCT example, we also found genes with large weights such as FGFR4 and MIC2 that were specific indicators of one cancer type. These genes were also confirmed independently [2, 4, 6]. This suggests that genes with greater weights represent "signature genes" for classifying different cancer types.

Using gene expression profiles as a clinical tool requires reproducibility and reliability. Measuring only a small number of genes that are essential for an accurate diagnosis would in practice simplify the assay and reduce the cost. Therefore, identifying a reliable and consistent list of genes such as the "signature genes" is important. It is interesting therefore that considerable variation in the gene lists was found both within our study and between other studies that examined the same data sets [1, 4, 5, 3]. The latter could be due to the different objective functions that different methods tried to optimize. In our study, we noticed that within our hundreds of runs, although some genes were always assigned large weights, other genes were assigned variable weights. However, despite the variability in gene lists, the classifiers had similar or equal classification accuracy. Mathematically, the stochastic searching algorithm in this work guaranteed a local optimal solution. Each learned classifier was an optimal solution to the classification task. The learned classifiers achieved similar accuracy with different sets of features. Theoretically these sets of features did not have to overlap. The biological relevance of these gene lists is unknown, but the similar classification accuracy suggests that these different sets may be functionally related. By looking into the pathways that those genes are involved, it could reveal the functional relationship among these genes.

So far, two benchmark data sets were tested in this current work with the maximum feature number less than 10000. Future work is focused on evaluating the performance of the proposed method on much larger data sets in comparison to other state-of-art methods, such as SVM and RF.

5. ACKNOWLEDGMENTS

This work was supported in part by NCI grant CA75056 and a gift from the C.B. Wang Foundation to the Center for Computational Genomics, Case Western Reserve University. We thank the anonymous reviewers for their insightful comments.

6. REFERENCES

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439): 531–537, 1999.

[2] J. Khan, et. al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[3] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene Selection for cancer classification using support vector machines. *Machine Learning*, 46, 389–422, 2002.

[4] R. Tibshirani, T. Hastie and G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99 (10): 6567–6572, 2002.

[5] A.V. Antonov, I.V. Tetko, M.T. Mader, J. Budczies and H.W. Mewes, Optimization moddels for cancer classificaiton: extracting ene interaction information from micrioarray expression data. *Bioinformatics*, Vol. 20, No.5, 644–652, 2004.

[6] H. Kovar et. al., Overexpression of the pseudoautosomal gene MIC2 in Ewing's sarcoma and peripheral primitive neuroectodermal tumor. *Oncogene*, 5, 1067–1070, 1990.

[7] Y. Yang and X. Liu, A re-examination of text categorization methods. *SIGIR'99: Proc. of 22nd annual inter. ACM SIGIR conf. on Research and development in information retrieval*, 42–49, 1999.

[8] Y. Freund and R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 55:119–139, 1997.

[9] A. Gammerman, V. Vovk, and V. Vapnik, Learning by transduction. In *Proc. of 14th Conf. on Uncertainty in Artificial Intelligence*, 148–155, 1999.

[10] V. Volodya, A. Gammerman and C. Saunders, Machine-learning applications of algorithmic randomness. In *Proc. of 16th Int. Conf. on Machine Learning*, 444–453, 1999.

[11] C. Saunders, A. Gammerman, V. Vovk, Transduction with confidence and credibility. In *Proc. of IJCAI'99*, 722–726, 1999.

[12] C. Saunders, A. Gammerman, V. Vovk, Computationally effecient transductive machines. Algorithmic Learning Theory, 11th International Conference, ALT 2000, Sydney, Australia, December 2000, Proceedings, 325–333, 2000.

[13] I. Kononenko, Estimating Attributes: Analysis and Extensions of RELIEF, In *European Conference on Machine Learning*, 171–182, 1994.

[14] L. Breiman, Random forests. *Machine Learning*, 45:5–32, 1999.

[15] F. Li, J. Kosecka and H. Wechsler. Strangeness based feature selection for part based recognition. *CVPRW*, 22–22, 2006.