Ensembles of Nearest Neighbours for Cancer Classification Using Gene Expression Data

Oleg Okun[1] and Helen Priisalu[2]

[1] University of Oulu, Oulu 90014, Finland,
[2] Tallinn University of Technology, Tallinn 19086, Estonia

Abstract. It is known that an ensemble of classifiers can outperform a single best classifier if classifiers in the ensemble are sufficiently diverse (i.e., their errors are as much uncorrelated as possible) and accurate. We study ensembles of nearest neighbours for cancer classification based on gene expression data. Such ensembles have been rarely used, because the traditional ensemble methods such as bagging and boosting are unable to inject diversity into nearest neighbour classifiers. To alleviate this problem, feature selection prior to classification is done. After that, diversity-based ensemble pruning before classifier combination can lead to an ensemble outperforming a single best nearest neighbour. However, such a result cannot always hold as demonstrated by experiments with two gene expression datasets. The reason seems to lie in dataset complexity. We show that if a dataset is easy for a single nearest neighbour to accurately classify, then an ensemble of nearest neighbours is hardly capable to get it better. In contrast, if a dataset is difficult according to certain complexity measures, then an ensemble of nearest neighbours is superior in performance to a single best nearest neighbour.

1 Introduction

Cancer prediction using gene expression data is one of the topics of intensive research in bioinformatics. It is based on the assumption that expression levels of genes in healthy and diseased cells are different so that it is possible to discriminate between cancer and normal state. However, the existing gene expression datasets bring a challenge for automatic discrimination: high dimensionality compared to the number of cases collected from patients.

To mitigate the negative effect of high data dimensionality, feature selection is usually recommended before classification. In the case of gene expression datasets, features are gene expressions so that feature selection is synonymous to gene selection. As follows from its name, feature selection involves finding a subset of the original features so that a classifier built with this subset would perform better than a classifier built from the entire set of features.

Given n genes expressed in N cases, the task is to select a small subset of all genes so that 1) the cases are accurately classified into two classes (normal and cancer) and 2) the selected subset consists of genes biologically relevant to

cancer. That is the task is twofold, because it involves both machine learning and biological aspects.

However, different feature selection methods will often select different subsets of features from the same dataset so that relevant features vary. To address this issue, it is often better to combine results of several feature selection methods. This strategy is based on the intuition that features that are consistently selected by various feature selection methods tend to be more relevant for the problem in question. Because feature selection precedes classification, it would be natural to combine these two operations together under the name 'base classifier', which we will use in this paper. As a result, we can now talk about ensembles of classifiers.

A typical ensemble fuses the predictions (either class labels or probability estimates) of a set of classifiers, thus forming a single class membership for the test cases. The objective of applying ensembles is to achieve superior performance over a single best classifier. One possibility to do so is to combine both diverse and accurate classifiers [1]. Diversity means that different classifiers do not make errors on the same cases. One of the most typical ways to make classifiers diverse is to let each classifier to work only with a certain subset of the original features so that these subsets can overlap but they do not totally coincide. These subsets can be randomly sampled from the original set like in [2] or selected by a certain feature-ranking algorithm. We rule out random sampling, since it can easily lead to picking the features that are biologically irrelevant to cancer. Instead, we rely on statistically-driven feature ranks, which provide a better guarantee against selecting the features purely by chance.

There are several articles employing ensembles for cancer classification, where ensembles are built with neural networks [3] and decision trees [4–6]. In contrast to these works, we focus on ensembles of nearest neighbour classifiers. Despite of being one of the classifiers of choice for many real-world tasks, a nearest neighbour (NN) is relatively rarely applied in ensembles, especially regarding cancer classification based on gene expression data (exceptions are [7–9]), because several researchers showed that neither bagging nor boosting (two most popular ensemble techniques) does provide improvement compared to a single NN classifier [2, 10]. In our opinion, the previous approaches based on NN ensembles suffer from the principal drawback related to automatic ensemble construction: either all possible ensemble configurations have to be tried in order to select one of them [8] or this process involves stochastic search [9], which leads to an unstable solution. We propose to apply the diversity-based ensemble pruning algorithm [11] to quickly and automatically create an ensemble out of individual NN classifiers, given the number of classifiers to fuse. Class labels assigned to training data by individual classifiers are only sufficient and no re-training is needed when including NNs into an ensemble. The latter characteristic is especially valuable, since one does not have to worry about ensemble overfitting.

Besides the abovementioned reasons inspiring our interest to NN ensembles, we were motivated by the fact that a nearest neighbour demonstrated a highly competitive performance with respect to more sophisticated classifiers [6]. Hence, as noticed in [7], one should not always expect that an NN ensemble will out-

perform a single NN. Although in [7] some efforts were made to answer when a single best NN classifier can outperform an NN ensemble, the results are still far from being comprehensive. In this work, we continue to study this problem by concentrating on dataset complexity as measured in [12]. We argue that this complexity can be the reason of why NN ensembles cannot outperform a single best NN classifier. When a dataset is complex, e.g., it includes multiple types of cancer, an NN ensemble is superior to a single best NN classifier. However, if a dataset is rather easy for classification, a highly accurate NN can render efforts of an ensemble worthless. Together with ensemble pruning, dataset complexity constitutes the novelty of our work, since these two issues are often overlooked in the literature on cancer classification.

2 Methods

2.1 Feature Selection Methods

We chose the following methods employed in [13]: pairwise feature selection, greedy forward selection, and individual ranking. As follows from its name, pairwise feature selection takes into account interaction between pairs of features. Other two methods are one-feature-at-a-time methods. All the methods expect two-class data and belong to the filter model, i.e., they do not rely on a classifier when selecting features.

Method of Bø and Jonassen This method evaluates how well a pair of genes in combination distinguishes two classes. First, each pair is evaluated by computing the projections of the training data on the diagonal linear discriminant (DLD) axis (see [13] for details) when using only two genes constituting this pair. Then the two-sample t-statistic is computed on the projected points and assigned to a given pair as its score. Bø and Jonassen proposed two variants of gene selection based on pair scores, called 'all pairs' (exhaustive search) and 'greedy pairs' (greedy search).

The all-pairs variant targets all possible pairs of genes. First, all pairs are sorted in descending order of their score. After that, the pair with the highest pair score is selected and all other pairs containing either gene included in this pair are removed from the sorted list of pairs. Then the next highest-scoring pair is found from the remaining pairs and all other pairs containing either gene in this pair are removed from the list, and so on. This procedure terminates when the pre-specified number of genes is reached.

Since the all-pairs variant is computationally demanding, an alternative evaluating only a subset of all pairs is proposed (greedy pairs). The greedy-pairs variant first ranks all genes based on indivudual t-score. Next, the best gene g_i ranked by its t-score is selected. Among all other genes, the gene g_j that together with g_i maximises the pair t-score is found so that g_i and g_j form a pair. These two genes are then removed from the gene set and a search for the next pair is performed until the desired number of genes is selected.

Greedy Forward Selection It is the greedy feature selection method, incrementally adding genes one-by-one; hence the name 'forward selection'. Individual t-score is computed for each gene and genes are ranked according to their t-score.

The first gene to be selected is the one with the highest t-score. Next, other genes are added so that adding a gene maximises the two-sample t-score of a new subset. The t-score at step i is computed on the data points projected on the DLD axis (as done in [13]) using the expression values of the i selected genes. Greedy forward selection stops when collecting the desired number of genes.

Individual Ranking The methods above belong to gene subset ranking techniques because each time a group of genes gets a rank [14] in contrast to individual ranking which grades each gene in isolation from others. Compared to gene subset ranking, individual ranking looks handicapped since it ignores interactions among genes. However, it is the fastest method among the four, because it is the simplest one. Each gene gets indivudual t-score ranking this gene among others. Gene selection is straightforward: a user picks the required number of top ranked genes.

2.2 Weighted k-Nearest Neighbour Classifiers

In our approach, classification follows feature selection. As classifiers, we employ the weighted k-nearest neighbour (k-NN) ($k = 1, 3, 5$), with weights assigned to the original genes as described in the experimental protocol[3]. It is known that a k-NN classifier is sensitive to irrelevant features [15]; hence, gene selection prior to classification will help to filter out irrelevant genes.

A k-NN classifier consists of a training set and a similarity (or distance) metric. To classify a new, unseen case, it computes distances to all cases in the training set and finds k nearest cases to the test case. The most common class of these k cases is assigned to the new case.

Given two n-dimensional cases x and y, the weighted distance between them is determined by the following formula:

$$\sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2} = \sqrt{\sum_{i=1}^{n} (\sqrt{w_i} x_i - \sqrt{w_i} y_i)^2} \ ,$$

where w_i is the weight of the ith gene. Thus, cases have to be weighted before searching for nearest neighbours.

2.3 Ensemble Pruning Before Classifier Combination

It is imperative that in order for an ensemble to outperform a single best classifier, the former must be composed of diverse classifiers that are also sufficiently

[3] The ith weight is equal to the number of times the ith gene has been selected during cross-validation of feature selection.

accurate at the same time [1]. In many models injecting diversity into an ensemble, e.g., via random sampling of input data like in [10] or via random feature sampling like in [2], diversity is not explicitly measured and controlled. Hence, it is difficult to judge how diverse the classifiers and this uncertainty appeals to incorporate some *explicit* measure of diversity into selection of the classifiers.

Though all the base classifiers are typically combined into an ensemble, it is possible to select only certain classifiers for fusion. This operation is called ensemble pruning. We borrowed the diversity-based pruning algorithm from [11]. In [11], they defined diversity for a pair of classifiers D_i and D_j as the number of cases where D_i and D_j produce different predictions (class labels). The diversity-based algorithm is given in Figure 1, where $\mathcal{H}$ and $\mathcal{C}$ are sets containing the original and selected classifiers, respectively.

First, the most accurate classifier is included into $\mathcal{C}$. Next, in each round, the algorithm adds to the list of selected classifiers the classifier D_k that is most diverse to the classifiers chosen so far, i.e., the D_k that maximises the diversity S over $\mathcal{C} \cup D_k, \forall k \in \{1, 2, \ldots, |\mathcal{H}|\}$. The selection ends when the M most diverse classifiers are selected.

Let $\mathcal{C} = \emptyset$, $M :=$ maximum number of classifiers
For $i := 1, \ldots, \mid \mathcal{H} \mid -1$ **do**
  For $j := i, \ldots, \mid \mathcal{H} \mid$ **do**
    Let $z_{i,j} :=$number of cases where
      D_i and D_j give different predictions
Let $D' :=$ the classifier with the highest accuracy
$\mathcal{C} := \mathcal{C} \cup D'$, $\mathcal{H} := \mathcal{H} - D'$
For $i := 1, \ldots, M - 1$ **do**
  For $j := 1, \ldots, \mid \mathcal{H} \mid$ **do**
    Let $S_j := \sum_{k=1}^{|\mathcal{C}|} z_{j,k}$
  Let $D' :=$ the classifier from $\mathcal{H}$ with the highest S_j
  $\mathcal{C} := \mathcal{C} \cup D'$, $\mathcal{H} := \mathcal{H} - D'$

Fig. 1. Diversity-based ensemble pruning

2.4 Classifier Combination

We employ two classifier combination methods: majority vote and Naïve Bayes. Notation and formulae in this section are borrowed from [1].

Let us assume that there are L classifiers D_i, $i = 1, \ldots, L$ and c different classes. The label outputs of these classifiers compose c-dimensional binary vectors $[d_{i,1}, \ldots, d_{i,c}] \in \{0, 1\}^c$, $i = 1, \ldots, L$, where $d_{i,j} = 1$ if D_i labels the input $\mathbf{x}$ in class ω_j, and $d_{i,j} = 0$ otherwise.

Majority Vote The majority vote results in an ensemble decision for class ω_k if this class got more votes than any other class. Ties are randomly resolved.

Naïve Bayes Combination This combination method assumes that individual classifiers are mutually independent; hence the name 'naïve'. Let $s_1, \ldots, s_L$ be crisp class labels assigned to $\mathbf{x}$ by classifiers $D_1, \ldots, D_L$, respectively. The independence assumption means that the support for class ω_k is $\mu_k(\mathbf{x}) \propto P(\omega_k) \prod_{i=1}^{L} P(s_i \mid \omega_k)$, where $P(s_i \mid \omega_k)$ is the probability that D_i assigns $\mathbf{x}$ to class s_i, given that the true class is ω_k. For each classifier D_i, a $c \times c$ confusion matrix CM^i is calculated by applying D_i to the training data. The (k,s)th entry of this matrix, $cm^i_{k,s}$ is the number of cases whose true class label was ω_k, but they were classified in class ω_s by D_i. By N_k we denote the total number of cases from class ω_k. Taking $cm^i_{k,s_i}/N_k$ as an estimate of the probability $P(s_i \mid \omega_k)$, and N_k/N ($N = \sum_{j=1}^{c} N_j$) as an estimate of the prior probability for class ω_k, we obtain that $\mu_k(\mathbf{x}) \propto \frac{1}{N N_k^{L-1}} \prod_{i=1}^{L} cm^i_{k,s_i}$.

To avoid nullifying $\mu_k(\mathbf{x})$ regardless of other estimates when an estimate of $P(s_i \mid \omega_k) = 0$, the last formula can be rewritten as $\mu_k(\mathbf{x}) \propto \frac{N_k}{N} \left\{ \prod_{i=1}^{L} \frac{cm^i_{k,s_i} + 1/c}{N_k + 1} \right\}^B$, where B is a constant (we always set it to 1 in experiments).

2.5 Performance Evaluation

A Receiver Operating Characteristic or ROC is a plot of false positive rate (X-axis) versus true positive rate (Y-axis) of a binary classifier. It is commonly used for visualising and selecting classifiers based on their performance, because it seems that it is one of the best measures of predictive accuracy. The true positive rate (TPR) is defined as the ratio of the number of correctly classified positive cases to the total number of positive cases. The false positive rate (FPR) is defined as the ratio of incorrectly classified negative cases to the total number of negative cases.

The diagonal line $y = x$ corresponds to a classifier which predicts a class membership by randomly guessing it. Hence, all useful classifiers must have ROC curves above this line.

The ROC curve is a two-dimensional plot of classifier performance. To compare classifiers one typically prefers to work with a single scalar value. This value is called the Area Under Curve or AUC. It is calculated by adding the areas under the ROC curve between each pair of consecutive FPR values, using, for example, the trapezoidal rule. Because the AUC is a portion of the area of the unit square, its value will always lie between 0 and 1. Because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5 [16]. In fact, the better a classifier performs, the higher its AUC. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive case higher than a randomly chosen negative case [16].

3 Experimental Protocol

The following experimental protocol was utilised in our experiments.

Input: Training set containing gene expressions; the number of cross-validation folds, m; G feature selection methods; Q nearest neighbour classifiers; the maximum number of genes, n^g_{max}, to be selected; threshold thr to filter out rarely selected genes; the number of base classifiers, M, to combine into an ensemble; the number of top genes, n^g_i, $i = 1, \ldots, G \times Q$, to be used by each base classifier.

Step 1: Partition the training set into m non-overlapping and stratified folds.

Step 2: For each of the G feature selection methods, repeat the following operations m times: leave out one fold, select n^g_{max} genes using the remaining folds, and store indices of the selected genes in a special list. Thus, m lists of length n^g_{max} are formed for each feature selection method.

Step 3: For each of the L ($L = G \times Q$) base classifiers, do:

- pick top n^g_i ($n^g_i \leq n^g_{max}$) genes in each of the m lists created in the previous step by the jth feature selection method and combine genes from all lists together;
- compute the frequency of occurrence (weight), w, of each gene in the combined set;
- filter out rare genes as follows: if $w < thr$ for a certain gene, set its w to zero; otherwise keep its w unchanged;
- using the updated values of w and folds created in Step 1, classify the training set using m-fold cross-validation and the ℓth weighted classifier;
- construct the ROC curve and compute the AUC value for this classifier.

Step 4: Perform ensemble pruning, which reduces the number of base classifiers from L to M.

Step 5: Combine the M remaining classifiers into an ensemble by one of the non-trainable combination schemes.

Step 6: Combine all genes used by the ensemble into one set. Compute the frequency of occurrence of each gene. Rank genes according to the frequency of occurrence.

Step 7: Construct the ROC curve and compute the AUC value for the resulting ensemble.

Otput: ROC curves and AUC values for individual classifiers and their ensemble, sets of genes used to cross-validate each classifier, a ranked list of genes created in Step 6.

Gene selection using the jth feature selection method followed by classification by the ℓth classifier is termed a base classifier. A combination of four feature selection methods ($G = 4$) and three NNs ($Q = 3$) produces twelve base classifiers ($L = 12$). Given L base classifiers, gene selection has to be done only G instead of L times, since each $n^g_i \leq n^g_{max}$. We experimented with two ways to set values of n^g_i's: 1) all n^g_i's are equal to a certain constant (we used 5, 10, 20, 30, 40, 50) and 2) each n^g_i is set to a random number between 1 and n^g_{max} ($n^g_{max} = 50$).

Genes are selected m times by each feature selection method, where m is the number of folds ($m = 5$). However, in general, m lists of genes differ. Combining all lists produces a single set of genes, which facilitates further analysis and biological interpretation. But not every gene in the combined set is important. For instance, rarely occurring genes were likely selected by chance and can be therefore safely eliminated. In contrast, genes that were often selected are considered to be relevant for classification. To filter out rare genes, we applied a simple thresholding technique: if a certain gene appeared in the combined set less than thr times, where thr is a fraction of m ($thr = 2$ in this paper), this gene was treated as noise and its weight w was set to zero. Besides nullifying the influence of rare genes, thresholding helps to prevent overfitting of the NN classifiers to noise. Also notice that a set of genes used to cross-validate classifiers is different to a certain degree from any of the m subsets of genes found during cross-validation of feature selection. Besides, feature selection does not employ a NN classifier to select genes, which also attributes to mitigate a bias of the reported results. Finally, since NNs do not require training, no dramatical bias can be introduced in the whole process of feature selection and classification.

Having predictions (either class labels or probability estimates[4] or both) of individual base classifiers, it is possible to combine them in order to achieve better performance. Sometimes, certain classifiers can be left out, i.e., M out of L classifiers are sufficient to obtain superior results over a single best NN classifier. The ensemble pruning technique we applied is based on diversity in predictions of individual classifiers. It only requires class labels of the individual classifiers and does not involve any training.

Finally, the M ($M = 3, 5, 7, 9, 11, 12$) classifiers that are left after ensemble pruning are fused by one of the known combination schemes. In this work, we utilise two of them: majority vote and Naïve Bayes combination with a correction for zeroes [1]. Both schemes do not require extra training. Thus, they do not introduce a bias to our results. Because we need probabilistic outputs in order to generate ROC curves and to compute AUC values for ensembles, we approximate probabilities as follows. For majority vote, it is the ratio of votes for a certain class to the total number of votes, while for Naïve Bayes combination, it is the class support μ. Cancer cases are positives whereas normal cases are negatives.

Genes used by individual base classifiers are combined together and for each gene the number of base classifiers using this gene is counted. Based on this characteristic, all genes can be ranked in order to guide a biological verification towards few top ranked genes, because genes frequently used by many classifiers in the ensemble can be highly relevant for future clinical studies.

4 Datasets

Two datasets were chosen for experiments.

[4] Laplace correction for k-NN classifiers [1] is used. Probabilistic outputs are then employed to build ROC curves and to compute AUC values.

The first dataset is produced according to SAGE technology. SAGE stands for Serial Analysis of Gene Expression[17]. This is technology alternative to microarrays (cDNAs and oligonucleotides). Though SAGE was originally conceived for use in cancer studies, there is not much research using SAGE datasets regarding ensembles of classifiers (to our best knowledge, this is the first research on NN ensembles based on SAGE data).

In this dataset[18], there are expressions of 822 genes in 74 cases (24 cases are normal while 50 cases are cancerous) [19]. Unlike many other datasets with one or few types of cancer, it contains 9 different types of cancer. We decided to ignore the difference between cancer types and to treat all cancerous cases as belonging to a single class. No preprocessing was done.

The second dataset is a microarray (oligonucleotide) dataset [20] introduced in [21]. It contains expressions of 2000 genes for 62 cases (22 normal and 40 colon tumour cases). Preprocessing includes the logarithmic transformation to base 10, followed by normalisation to zero mean and unit variance as usually done with this dataset.

We will use the names 'SAGE' and 'Colon' in order to identify these datasets.

5 Experimental Results

We used the AUC value for performance evaluation of both single classifiers and ensembles. In the first series of experiments, we fixed the number of genes to be selected and used with each base classifier in the ensemble to a certain constant C. Hence, this number was the same for all classifiers in the ensemble. Since each feature selection method ranks all genes, it is natural to select top C genes according to the ranks assigned by each method. Thus, feature selection helps to concentrate on a certain subset of all genes. The following values of C were tried: 5, 10, 20, 30, 40, 50. For ensembles rare genes were eliminated at $thr = 2$.

Regarding SAGE, one can notice that NN ensembles created with both combination schemes often outperformed a single best NN classifier, especially when using 20-30 genes per base classifier. Moreover, it was sufficient to fuse 5-9 NN classifiers in order to achieve the superior performance over the ensemble including all available classifiers. This testifies to the usefulness of ensemble pruning prior to classifier combination. In addition, pruning limits the number of genes to be further analysed, because the fewer classifiers, the fewer subsets of selected genes. For Colon, NN ensembles did not demonstrate such superiority, i.e., a single best NN turned out to be more accurate than an NN ensemble, regardless of the combination scheme. Regarding the absolute AUC values, the AUC values obtained on the SAGE data tend to be much lower than those of the Colon data.

The results above point to the difference in performance of NN ensembles on two datasets and that it should not always be expected that an NN ensemble will outperform a single best classifier. To explain these facts, we measured dataset complexity by means of a number of characteristics borrowed from [12]. It turned out that according to the majority of them, SAGE is more complex than Colon. It is not surprising since the former comprises 9 types of cancer

while the latter contains only one. The reason for a higher complexity of SAGE is highly overlapped classes. That is why, NN ensembles achieved better results on the less complex Colon data. Thus, dataset complexity analysis is very useful because one can roughly estimate the expected performance on the dataset under study before classifying it. Based on the dataset complexity, it is also possible to explain why a single NN sometimes performs better than an NN ensemble. If a dataset is easy to classify, a single best NN classifier (SBC) may be able to do so with high accuracy. In this case, it is likely that the other classifiers are greatly inferior to SBC. Hence, combining one good classifier and several bad classifiers will drown out the vote of SBC in the votes of incorrect classifiers. As a result, the whole ensemble will be unable to outperform an SBC. In contrast, when a dataset is difficult to classify, a SBC is a weak classifier and so are the other classifiers. An ensemble of such weak classifiers can therefore improve on the SBC result [1, 22].

To confirm our hypothesis about the link between dataset complexity and classification performance, we carried out a broader experiment, this time with the variable number of genes for base classifiers in the ensemble. Thus, the number of top genes varied from classifier to classifier. The variable number of top genes allowed us to consider a more general case. Individual NNs and NN ensembles of different structures were created 500 times. The AUC values of both SBCs and ensembles were collected and based on this information, statistical tests were carried out. We tested the hypothesis whether an NN ensemble can outperform an SBC. The Kruskal-Wallis nonparametric one-way Analysis of Variance (ANOVA) was done, followed by the multiple comparisons with Scheffé's S adjustment for the significance level α ($\alpha = 0.01$ or $\alpha = 0.05$). See[23] for details of these tests. Based on the results of these tests, NN ensembles indeed fared better than SBCs for SAGE but not for Colon, where SBCs were superior to ensembles. Thus, statistical tests on a larger sample confirmed our findings reported above when all base classifiers in the ensemble worked with the same number of genes defined by C.

Besides performance analysis, we were interested to see how often different feature selection methods were coupled with SBCs (and therefore the most accurate classifiers in each ensemble). The information on this subject is collected for 500 ensembles per combination scheme. It turned out that for SAGE dataset, greedy forward selection emerged as a strong leader (56% of cases). For Colon dataset, the all-pairs Bø-Jonassen algorithm was clearly dominant (85% of cases). In contrast, individual ranking rarely led to the best performance of an individual NN (7.1% and 1.7% of cases for SAGE and Colon data, respectively). Greedy forward selection and all-pairs Bø-Jonassen perform a broader search than individual ranking. Hence, they are capable of finding a better solution, but at a higher computational cost.

5.1 Biological Relevance of Selected Genes

Due to space limitations, we consider only one case: SAGE dataset and rare gene filtering at $thr = 2$. Colon dataset has been already studied in many works

while SAGE dataset is rather new. Besides, it contains multiple types of cancer so that it is more interesting for analysis of gene relevance.

In focusing on genes for biological interpretation, we noticed another advantage of classifier ensembles. Counting the number of times when a certain gene (among the selected genes) was utilised by the base classifiers can indicate the importance or relevance of this gene. Thus, this count serves as a gene rank. Combining results from several ensembles can provide even better (more stable) ranking. For instance, when applied to the combined list of genes generated for six values of C, the top 4 genes are:

Gene 721 TGAGTGGTCA ESTs, Highly similar to HYPOTHETICAL 13.6 KD PROTEIN in NUP170-ILS1 INTERGENIC REGION [Saccharomyces cerevisiae].

Gene 409 GAGTGGGGGC ESTs ('EST' means expressed sequence tag), Highly similar to LYSOSOMAL PRO-X CARBOXYPEPTIDASE PRECURSOR [Homo sapiens].

Gene 596 GTGCATCCCG Casein kinase 2, beta polypeptide.

Gene 629 GTTCTCCCAC ESTs, Highly similar to PROTEIN TRANSPORT PROTEIN SEC61 ALPHA SUBUNIT [Canis familiaris].

Our first expectation was that since different types of cancer are present in the dataset, there should be genes (among the top selected) that can be found in many human tissues. It turned to be true for all four genes. We consulted the paper [24] and Table 6 [25] supplementing that paper and containing the list of *ubiquitously* expressed genes (i.e., these genes can be found in many human organs) as well as gene descriptions found on the Web.

Gene 721 can be found in many human tissues such as brain, breast, colon, heart, kidney, prostate, skin, etc. However, the link of gene 721 to cancer is obscure and it can be that it is just ubiquitous gene. We found one study [26] pointing to the high expression level of this gene in chromophobe carcinoma [27], which is one of the most common histological subtypes of renal cell carcinoma (RCC), which, in its turn, is one of the 10 most frequent malignancies in Western societies.

Gene 409 may be related to essential hypertension [28]. There is evidence that hypertension may be related to cancer [29]: for example, high blood pressure in combination with smoking can cause lung cancer [30]. However, this hypothesis is still disputed [31] for other types of cancer: in particular, it may be so that some antihypertensive drugs are associated with the risk of cancer, but a precise mechanism of such association is yet unknown. Currently, all that researchers seem to agree is that cancer and hypertension share some common pathological mediators, and through these, hypertension might influence on malignant growth of cancer. LYSOSOMAL PRO-X CARBOXYPEPTIDASE PRESURSOR was also found among top 50 marker genes to classify acute lymphoblastic/mieloid leukemia (ALL/AML) in [32].

Gene 596 can be connected to cancer [33, 34]. This gene encodes the beta subunit of casein kinase II, a ubiquitous protein kinase which regulates metabolic pathways, signal transduction, transcription, translation, and replication [35].

Thus, it is a very important gene. It had been observed that protein kinase CK2 is elevated in various cancers [36]. In addition to cancers, CK2 is found to interact with the human immunodeficiency virus type 1 (HIV-1) proteins and the human herpes simplex virus type 1 (HSV-1) protein EB63 [35].

Gene 629 is highly similar to the canine gene. Its cellular function can be protein synthesis and transport. In medicine, the human safety often demands prior tests involving animals, including dogs. By observing effects of such clinical tests, a greater understanding of various diseases and their potential cure can be achieved, though extrapolation from dogs to humans should not be always straightforward, because, for example, metabolic and toxic outcomes differ across species. Recently, in [37], it was found that Sec61a, involved in protein secretion, strikingly responds to inflammatory stimuli. Another study [38] also showed that Gene 629 is involved in the inflammatory process in mice. It is estimated that inflammation contributes to the development of at least 15% of all cancers [39]. Therefore Gene 629 may be indirectly linked to cancer.

Besides top genes, interesting information can be digged out in other genes, too. For example, let us consider the following two genes:

Gene 96 AGCCACCACG Human mRNA for KIAA0149 gene, complete cds.
Gene 594 GTGATCTCCG ESTs.

Gene KIAA0159 is a scavenger receptor expressed by endothelial cells that are involved in the control of blood pressure, blood clotting, atherosclerosis, and inflammation. These cells are located on the interior surface of blood vessels and form a bridge between blood in the lumen and the rest of the vessel wall. Scavenger receptors have been implicated in the pathogenesis of atherosclerosis, and there is enough evidence to associate atherosclerosis with cancer [40, 41] (also Google provides more than 25000 links to the phrase "atherosclerosis and cancer"). In other words, both diseases may represent variants of a common disease entity.

Gene 594 was found in 83 mRNA-source sequences clustered in two large groups [42]. The description of SAGE libraries points to connections of this gene to various cancers and Gaucher disease. Researchers hypothesised that patients with Gaucher disease might be at a higher risk of such malignant disorders as cancer [43, 44]. However, because Gaucher disease is relatively rare, it is diffucult to study the links between it and cancers, using large samples. Nevertheless, there is quite strong indication that Gaucher disease and hematological cancers (lymphoma, myeloma) are related [44, 45].

6 Conclusion

In this paper, we studied cancer classification using gene expression data. Two real-world datasets different in the complexity and the number of cancer types were utilised in experiments. We demonstrated 1) the importance of feature selection in making the ensembles of NNs diverse, 2) if dataset complexity is

high, e.g., a dataset is composed of cases belonging to multiple types of cancer, small NN ensembles composed of 5-9 classifiers can outperform a single best NN classifier as attested by the AUC value, 3) NN ensembles can benefit from diversity-based pruning prior to classifier combination, 4) using measures of dataset complexity can help to evaluate the expected performance of individual NN classifiers and NN ensembles on this dataset. From the biological point of view, employing the NN ensembles helps to concentrate on small groups of genes thanks to two mechanisms: filtering out rarely selected genes and those genes that were only utilised by few base classifiers in the ensemble.

References

1. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Inc.,, Hoboken, NJ (2004)
2. Bay, S.: Nearest neighbor classification from multiple feature sets. Intelligent Data Analysis **3**(3) (1999) 191–209
3. Liu, B., Cui, Q., Jiang, T., Ma, S.: A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics **5** (2004) 136
4. Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics **2**(Suppl 3) (2003) 75–83
5. Long, P.M., Vega, V.B.: Boosting and microarray data. Machine Learning **52**(1-2) (2003) 31–44
6. Dudoit, S., Fridlyand, J.: Classification in microarray experiments. In Speed, T., ed.: Statistical Analysis of Gene Expression Microarray Data. Interdisciplinary Statistics. Chapman and Hall/CRC, Boca Raton, Florida (2003) 93–158
7. Paik, M., Yang, Y.: Combining nearest neighbor classifiers versus cross-validation selection. Statistical applications in genetics and molecular biology **3**(1) (2004) Article 12
8. Cho, S.B., Won, H.H.: Data mining for gene expression profiles from dna microarray. Int J of Software Engineering and Knowledge Engineering **13**(6) (2003) 593–608
9. Park, C., Cho, S.B.: Evolutionary computation for optimal ensemble classifier in lymphoma cancer classification. In Carbonell, J.G., Siekmann, J., eds.: Proceedings of the Fourteenth International Symposium on Methodologies for Intelligent Systems, Maebashi City, Japan. (2003) 521–530
10. Breiman, L.: Bagging predictors. Machine Learning **24**(2) (1996) 123–140
11. Prodromidis, A.L., Stolfo, S., Chan, P.K.: Pruning classifiers in a distributed meta-learning system. In: Proceedings of the First Panhellenic Conference on New Information Technologies: 8-10 October 1998; Athens, Greece. (1998) 151–160
12. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans on Pattern Analysis and Machine Intelligence **24**(3) (2002) 289–300
13. Bø, T.H., Jonassen, I.: New feature subset selection procedures for classification of expression profiles. Genome Biology **3**(4) (2002) 0017.1–0017.11
14. Lu, Y., Han, J.: Cancer classification using gene expression data. Information Systems **28**(4) (2003) 243–268
15. Aha, D.W.: Generalizing from case studies: a case study. In Sleeman, D.H., Edwards, P., eds.: Proceedings of the Ninth International Workshop on Machine Learning: 1-3 July 1992; Aberdeen, Scotland, Morgan Kaufmann (1992) 1–10

16. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**(8) (2006) 861–874
17. Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W.: Serial analysis of gene expression. Science **270**(5235) (1995) 484–487
18. (http://lisp.vse.cz/challenge/ecmlpkdd2004/)
19. Gandrillon, O.: Guide to the gene expression data. In Berka, P., Crémilleux, B., eds.: Proceedings of the ECML/PKDD Discovery Challenge Workshop: 20 September; Pisa, Italy. (2004) 116–120
20. (http://microarray.princeton.edu/oncology/affydata/index.html)
21. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci **96**(12) (1999) 6745–6750
22. Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32
23. Zar, J.H.: Biostatistical Analysis. Prentice Hall, Inc.,, Upper Saddle River, NJ (1999)
24. Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., Cook, B.P., Dufault, M.R., Ferguson, A.T., Gao, Y., He, T.C., Hermeking, H., Hiraldo, S.K., Hwang, P.M., Lopez, M.A., Luderer, H.F., Mathews, B., Petroziello, J.M., Polyak, K., Zawel, L., Zhang, W., Zhang, X., Zhou, W., Haluska, F.G., Jen, J., Sukumar, S., Landes, G.M., Riggins, G.J., Vogelstein, B., Kinzler, K.W.: Analysis of human transcriptomes. Nature Genet **23** (1999) 387–388
25. (http://www.nature.com/ng/journal/v23/n4/extref/ng1299-387b S1.pdf)
26. Boer, J.M., Huber, W.K., Sültmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Füzesi, L., Vingron, M., Poustka, A.: Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. Genome Research **11** (2001) 1861–1870
27. (http://www.dkfz heidelberg.de/abt0840/whuber/rcc/clear_chromo_genes.txt)
28. (http://harvester.embl.de/harvester/P427/P42785.htm)
29. Hamet, P.: Cancer and hypertension: a potential for crosstalk? J Hypernens **15** (1997) 1573–1577
30. Lingren, A., Pukkala, E., Nissinen, A., Tuomilehto, J.: Blood pressure, smoking, and the incidence of lung cancer in hypertensive men in North Karelia in Finland. Am J Epidemiol **158** (2003) 442–447
31. Silva, D.G., Schönbach, C., Brusic, V., Socha, L.A., Nagashima, T., Petrovsky, N.: Identification of "pathologs" (disease-related genes) from the RIKEN mouse cDNA dataset using human curation plus FACTS, a new biological information extraction system. BMC Genomics **5** (2004)
32. Chow, M.L., Moler, E.J., Mian, I.S.: Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. Physiol. Genomics **5** (2001) 99–111
33. Faust, R.A., Gapany, M., Tristani, P., Davis, A., Adams, G.L., Ahmed, K.: Elevated protein kinase CK2 activity in chromatin of head and neck tumors: association with malignant transformation. Cancer Letters **101** (1996) 31–35
34. Wang, H., Davis, A., Yu, S., Ahmed, K.: Response of cancer cells to molecular interruption of the CK2 signal. Molecular and Cellular Biochemistry **227** (2001) 167–174
35. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=1460)

36. Guerra, B., Issinger, O.G.: Protein kinase CK2 and its role in cellular proliferation, development and pathology. Electrophoresis **20** (1999) 391–408
37. Higgins, M.A., Berridge, B.R., Mills, B.J., Schultze, A.E., Gao, H., Searfoss, G.H., Baker, T.K., Ryan, T.P.: Gene expression analysis of the acute phase response using a canine microarray. Toxicological Sciences **74** (2003) 470–484
38. Ibrahim, S.M., Koczan, D., Thiesen, H.J.: Gene-expression profile of collagen-induced arthritis. J Autoimmunity **18**(2) (2002) 159–167
39. Marx, J.: Inflammation and cancer: the link grows stronger. Science **306**(5698) (2004) 966–968
40. Li, J.J., Gao, R.L.: Should atherosclerosis be considered a cancer of the vascular wall? Medical Hypotheses **64**(4) (2005) 694–698
41. Ross, J.S., Stagliano, N.E., Donovan, M.J., Breitbart, R.E., Ginsburg, G.S.: Atherosclerosis and cancer. Annals of the New York Acad Sci **947** (2001) 271–293
42. (http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd=tagsearch&anchor= NLAIII&org=Hs&tag=GTGATCTCCG)
43. Rosenbloom, B.E., Weinreb, N.J., Zimran, A., Kacena, K.A., Charrow, J., Ward, E.: Gaucher disease and cancer incidence: a study from the gaucher registry. Blood **105**(12) (2005) 4569–4572
44. Shiran, A., Brenner, B., Laor, A., Tatarsky, I.: Increased risk of cancer in patients with Gaucher disease. Cancer **72**(1) (1993) 219–224
45. deMagalhaes Silverman, M.: Cancer and gaucher disease. Gaucher Clinical Perspectives **4**(2) (1996)