Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Ranked selection of nearest discriminating features

Alex Pappachen James[1*] and Sima Dimitrijev[2]

*Correspondence:
a.james@iiitmk.ac.in
[1]School of Computer Science, Indian Institute of Information Technology and Management (IIITM) - Trivandrum, Kerala, India
Full list of author information is available at the end of the article

Abstract

Background: Feature selection techniques use a search-criteria driven approach for ranked feature subset selection. Often, selecting an optimal subset of ranked features using the existing methods is intractable for high dimensional gene data classification problems.

Methods: In this paper, an approach based on the individual ability of the features to discriminate between different classes is proposed. The area of overlap measure between feature to feature inter-class and intra-class distance distributions is used to measure the discriminatory ability of each feature. Features with area of overlap below a specified threshold is selected to form the subset.

Results: The reported method achieves higher classification accuracies with fewer numbers of features for high-dimensional micro-array gene classification problems. Experiments done on CLL-SUB-111, SMK-CAN-187, GLI-85, GLA-BRA-180 and TOX-171 databases resulted in an accuracy of 74.9±2.6, 71.2±1.7, 88.3±2.9, 68.4±5.1, and 69.6±4.4, with the corresponding selected number of features being 1, 1, 3, 37, and 89 respectively.

Conclusions: The area of overlap between the inter-class and intra-class distances is demonstrated as a useful technique for selection of most discriminative ranked features. Improved classification accuracy is obtained by relevant selection of most discriminative features using the proposed method.

Background

Many of the contemporary databases used in data classification research [1-10] uses considerably large number of data points to represent an object sample. High dimensional feature vectors that result from these samples often contain intra-class natural variability reflected as noise and irrelevant information [11,12]. The noise in feature vectors occurs due to inaccurate feature measurements, whereas irrelevancy of a feature depends on the natural variability and the redundancy within the feature vector. Further, relevance of a feature is application dependent. For example, consider a hypothetical image consisting of image regions that correspond to faces and some other objects. When using this image in a face recognition application, the relevant pixels in the image are in the face regions while the pixels in the remaining regions are irrelevant. In addition, face regions themselves can have irrelevant information due to intra-class variability such as occlusions, facial expressions, illumination changes, and pose changes. Natural variability that

occurs in high dimensional data has significant impact on lowering the performance of all pattern recognition methods. To improve the recognition performance of classification techniques methods, in the recent past, most of the effort has been to compensate or remove intra-class natural variability from the data samples through various feature processing methods.

Dimensionality reduction [13-15] and feature selection [6-9] are two types of feature processing techniques that are used to automatically improve the quality of data by removing irrelevant information. Dimensionality reduction methods are popular because they achieve the purpose of reducing the number of features and noise in a feature vector with the mathematical convenience of feature transformations and projections. However, the assumption of correlations between the features in the data is a core aspect of dimensionality reduction methods that can result in inaccurate feature descriptions. Further, irrelevant information from the original data is not always possible to remove in a dimensionality reduction approach. Improving the quality of resulting features using linear and more recently non-linear dimensionality reduction methods has consistently been a field of intense research and debate in the recent past [13]. An alternative to dimensionality reduction approach, instead of trying to improve overall feature quality, feature selection tries to remove irrelevant features from the high dimensional feature vector thereby improving the performance of classification systems. Feature selection have been an intense field of study in the recent years, gaining importance in parallel with the dimensionality reduction methods. Feature selection provides an advantage over dimensionality reduction methods because of its ability to distinguish and select the best available features in a data set [6-10,16]. This means that feature selection methods can be applied to both the original feature vectors and to the feature vectors that result from the application of dimensionality reduction methods. From this point of view, feature selection can be considered as an essential component required for developing high performance pattern classification systems that use high dimensional data [1-3,17]. Since higher dimensional feature vectors contain several irrelevant features that reduce the performance of pattern recognition methods, feature selection by itself can be used in most of the modern data classification methods to combat the issues resulting from the curse of high dimensionality [18,19].

Feature selection problems revolve around the correct selection of feature subset. In a search-criteria approach to feature selection, feature selection is reduced to a search problem that detects an optimal feature subset based on the selected criteria. Exhaustive search ensures optimal solution, however, with increase in dimensionality such a search is computationally prohibitive. In the present literature, there exists no other distinct way to optimally select the features without reducing classification performance.

The existing research in feature selection has been focused on excluding features that are determined as most redundant using various search strategies and criteria assessment techniques[20-25]. In this paper, we propose a new method for feature selection based solely on individual feature discriminatory ability as an alternative to the existing search and criteria driven feature selection methods. The discriminatory ability of each feature is measured by the area of overlap between inter-class and intra-class distances that are obtained from feature to feature comparisons. Experimental results of a classification task based on microarray and image databases validate the effectiveness and accuracy of features obtained by our feature selection method.

Related work

Feature selection methods can be classified in three broad categories: *filter model* [26,27], *wrapper model* [28,29] and *hybrid and embedded model* [30,31]. In order to evaluate and select features, filter models exclusively use characteristics about the data, warper models uses mining algorithms, and hybrid models combine the use of characteristics about the data with data-mining algorithms. In general, these feature selection methods consists of three steps: (1) feature subset generation, (2) evaluation, and (3) stopping criteria [32]. Subset generation process is used to arrive at a starting set of features using different types of forward, backward or bidirectional search methods . Some of the most common techniques employed are complete search such as *branch and bound* [33] and *beam search* [34], sequential search such as *sequential forward selection*, *sequential backward elimination*, *bidirectional selection* [35], and random search such as *random-start hill-climbing* and *simulated annealing*[34]. The generated subset is evaluated for goodness using either an independent or a dependent criterion. Independent criterion is generally used in filter model, the popular ones are distance, dependency and consistency measures [35-37]. The dependent criteria is generally used in wrapper model requiring tuning of data-mining algorithms. The wrapper models perform better, however are computationally expensive and less robust to parameter changes in data-mining algorithms [38-41]. The goodness of the subsets using a selection criteria is assessed against stopping criteria such as minimum number of features, optimal number of iterations and lower classification error rates.

It can be noted that in conventional feature selection methods, features or subset of features are selected based on the rank as obtained by evaluating features against a selection criterion such that redundancy of features in the training set is minimized. The best performing methods for classification that rely on data-mining strategies include feature relevance calculations to select features holistically [20-22]. However, data-mining based solutions result in features that tend to be sensitive to minor changes in training data. Further, an increase in dimensionality makes the data-mining algorithms computationally intensive and often require problem specific optimization techniques. Contrary to data-mining based solutions, criteria driven methods based on filter models are computationally less complex and are more robust to minor changes in training data [23-25]. In such methods, the accuracy of initial selection of subsets using exhaustive forward or backward search of the features [42] would significantly impact the accuracy of features obtained with a given feature selection criterion. In addition, as pointed out in [28] optimal selection of subsets is intractable and in some problems are NP-hard [43]. Further, variations in the nature of data from one database to another make the optimal selection of an objective function difficult and a high classification accuracy using selected features from such methods are not always guaranteed. Because of such deficiencies, hybrids of filter and wrapper models also reflect these problems at various levels of feature selection.

The determination of inter-feature dependency as described by filter models, and wrapper models lay the foundations of present day feature selection methods. These models arrive at features that are often tuned to suite a classifier using several machine learning strategies at selection or criteria assessment stage. Some of the recent approaches that attempt to improve the performance of the conventional feature selection methods use the ideas of neighborhood margins [44-46], and manifold regularization using SVMs [47]. However, similar to wrapper methods that uses specific mining techniques, these recent methods are computationally complex and require additional optimization methods to

speedup calculations. In addition, optimal performance of the selected features on classifiers are highly sensitive to minor changes in training data and tuning parameters. Due these reasons, the practical applicability and robustness of such methods on large sample high dimensional datasets are questionable.

Conventional feature selection methods apply multiple level processing on a given feature vector to find a subset of useful features for classification using several machine learning techniques and search strategies. The presented work on the contrary draws specific attention to select most discriminating features from a single step process of discriminating subset selection. As distinct from the general idea of optimizing feature subsets for classification oriented filter and warper models, here we focus on developing an approach to determine relevant features from a training set solely by calculating their individual inter-class discriminatory ability.

Discriminant feature selection based on nearest features

Although not popular in feature selection literature, perhaps the simplest way to understand discriminatory nature of feature in a training set with two classes can be by using a search using naive bayes classifier. A low probability of error of individual features as obtained using baysian classifier would indicate good discriminatory ability and asserts the usefulness of the feature.

A standard approach in feature selection literature is to directly apply training and selection criteria on the feature values. However, when natural variability in the data is high and number of training samples are less, even minor changes in feature values would introduce errors in the bayes probability calculations. Classification methods such as SVM on the other hand try to get around this problem by normalising the feature values and by parametric training of the classifiers against several possible changes in features values. In classifier studies, this essentially shifts the focus from feature values to distance values. Instead of directly optimising the classifier parameters based on feature values, the distance functions itself is trained and optimised.

Proposed method

In this work, we attempt to develop a technique of feature selection by using the new concept of distance probability distributions. This is a very different concept to that of filter methods that applies various criterion such as inter-feature distance, bayes error or correlation measures to determine set of features having low redundancy. Instead of complicating the feature selection process by different search and filter schemes to remove redundant features and to maintain relevant features, we focus our work in using all features that are most discriminative and useful for a classifier. Further, rather than looking at feature selection as a problem of finding inter-feature dependencies for reducing number of features, we treat each feature individually and arrive at features that would have the ability to contribute to classifiers performance improvement.

Suppose there are M classes in a training set having patterns with a set of J features, with ω_{ij} as class label for feature j, where $i \in \{1..M\}$ and $j \in \{1..J\}$. And let x_{jk} be a feature in the kth training pattern that can be used to calculate the inter-class and intra-class distance probability distributions. The intra-class distances y_j^a of the jth feature in a training set is equal to the distance $1 - e^{-|x_{jk} - x_{j\bar{k}}|}$, where $k \in \{1..K\}$,$\bar{k} \in \{1..K\}$ with $k \neq \bar{k}$ within a class in training set with K samples. The inter-class distances y_j^e of a

feature x_{jk} in a training set belonging to a class ω_{ij} is equal to the distance $1 - e^{-|x_{jk} - \bar{x}_j|}$, where $\bar{x}_j$ is a feature at j belonging to a sample in another class other than that of x_{jk}. We can represent the set of classes that does not belong to the class ω_{ij} as $\bar{\omega}_{ij}$. Then the intra-class distance probability distribution of feature j in class ω_{ij} is $p(y_j^a|\omega_{ij})$ and the corresponding inter-class distance probability distribution is $p(y_j^e|\bar{\omega}_{ij})$. The area of overlap of these distributions can be seen as the probability of error of feature at j for a class label at i and represents the discriminatory ability of feature. Since, in practice we are dealing with samples in discrete form the probability density can be represented in discrete from with m bins, and the area of overlap $P_{(j|i)}$ can be represented as:

$$P_{(j|i)} = \frac{1}{2} \sum_{m=-\infty}^{y_0} p_m(y_j^a|\omega_{ij})dy + \frac{1}{2} \sum_{m=y_0}^{\infty} p_m(y_j^e|\bar{\omega}_{ij})dy \tag{1}$$

The relative area of overlap of feature among all the classes can be then found as:

$$\hat{P}_{(j|i)} = P_{(j|i)} - \min_i P_{(j|i)} \tag{2}$$

The minimum area of overlap for feature across different classes can be then calculated as a measure to establish the discriminatory ability of feature:

$$\hat{P}_j = 1 - \min_i \hat{P}_{(j|i)} \tag{3}$$

Taking the minimum value of $\hat{P}_{(j|i)}$ across different classes ensures that features that could discriminate well for any one of the class among many and such features can be considered as useful for classification. The features are ranked in descending order based on the value of $\hat{P}_j$, a value of 0 would force the feature to take a low rank while a value of 1 would force the feature to take top rank. Let R represent the set of $\hat{P}_j$, arranged in the order of their ranks, each rank representing feature or group of features. R set can be used to form a rank based probability distribution by normalising the $\hat{P}_j$.

It is well known that almost every other ranked distributions of empirical nature originating from realistic back end data follow a power law distribution. The top ranked features in a ranked distribution often retain most of the information. This effect is observed in different problems and applications, and has formed the basis of Winner-take-all and Pareto principles.

The ranked distribution is formed with $\bar{P}_r = \frac{\hat{P}_j}{\sum_{j=1}^{J} \hat{P}_j}$ represent the normalised value of $\hat{P}_j$ for the feature at j having a rank r. The cumulative ranked distribution c_j^r is obtained as:

$$c_r = \hat{P}_r + c_{r-1}, \text{ where } c_{-1} = 0 \tag{4}$$

The top ranked values of c_r can used to select the most discriminative set of features. Applying the winner-take-all principle, and in the lines of $20 - 80$ concept of rank-size distributions, it is logical to assume that the top ranked features would have maximum amount of discriminative information. The subset of features X having a size $L \in [1, J]$ from the ranked features can be selected based on a selection threshold θ.

$$x_j \in X \iff c_r \leq \theta \tag{5}$$

In other words, the features x_j corresponding to the ranks that fall below the cumulative area threshold θ is selected to form X with size L. The selection threshold θ for selecting the top ranked features is done using the proposed Def 1.

Definition 1. *The selection threshold θ is equal to the standard deviation σ of the distribution of c_j^r, where* $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_j^r - \frac{1}{N}\sum_{i=1}^{N} c_j^r)^2}$.

If each feature in X is uncorrelated and independent, the features within X will be very few or no be redundant features. The selection of X based on the discriminatory ability is sufficient to ensure good classification performance. However, in feature selection problem, there is a chance that the subset of discriminant feature would have very similar features, and such features become redundant in improving classification performance. Identifying the independence of discriminant features would ensure the detection of least redundant features. For two features, $\{x_r, x_{r+1}\}$, ranked in order of $\bar{P}_r$ and $\bar{P}_{r+1}$ values, let $p(x_r)$ and $p(x_{r+1})$ be the probability density functions, and $p(x_r, x_{r+1})$ be the joint probability density function, where $r \in [1, L]$ is the rank of a feature in X corresponds to an index j in the original feature space. Then the features are independent if it can be established that $p(x_r, x_{r+1}) = p(x_r)p(x_{r+1})$. This idea of independence testing is utilised in finding an independence score of a feature. The area score between the probability densities $p(x_r, x_{r+1})$ and $p(x_r)p(x_{r+1})$ in discrete domain is calculated as:

$$A_{r,r+1} = \frac{1}{2}\sum_{m=-\infty}^{x_0} p_m(x_r)p_m(x_{r+1})dx + \frac{1}{2}\sum_{m=x_0}^{\infty} p_m(x_r, x_{r+1})dx \tag{6}$$

The independence score I_r of feature x_r with respect to remaining $L-1$ features in X is determined as:

$$I_r = \frac{1}{L-1}\sum_{r=1}^{L-1} A_{r,r+1} \tag{7}$$

A value of $I_r = 1$ would indicate that x_r is an independent feature in X (or x_j in the feature set with jth feature in the original feature space corresponding to the rth rank feature in X), while a value of I_r would indicate that x_r is redundant and should be removed. The independence score I_r corresponding to the feature at j in the sample along with the discriminatory score $\hat{P}_j$ can be used to select the most independent set of discriminant features.

$$z_s = x_j \iff I_r\hat{P}_j \leq \epsilon \tag{8}$$

where the value of $\epsilon = 0.01$ is a small number, and z_s is the set of most relevant discriminative independent features x_j, with $s \leq J$.

These subset of top ranked features are considered as useful for classification. However, parameters and nature of decision boundary imposed by a specific classifier need to be considered before these features can be used for classification. Consider using a nearest neighbour classifier, then the relative importance of feature $z_s \in X$ can be rated based on the recognition performance of using individual feature z_s alone for classification. Assuming the independence of features, using a leave one out cross validation, the classification accuracy of sth feature and jth sample in training set with size J, and $l \in J$ is found by the identification of the class as:

$$w^* = \arg\min_{l, l\neq j} d(z_{sj}, z_{sl}) \tag{9}$$

The selected features z_s are ranked based on the total number of correct class identification w^* in descending order. The top ranked features represent the most discriminant features while the lower ranked ones are relatively of lower in class discriminatory ability when using a nearest neighbour classifier. Such a ranking of the features for a given classifier identifies itself as the best responding features for that classifier.

Results and discussion

The role of feature selection methods in a high dimensional pattern classification problem is to select the minimum number of features that maximize the recognition accuracy. In this section, we demonstrate how the newly proposed selection method performs this task on standard databases used for bench marking feature selection methods.

Advancements in measurement techniques and computing methodologies have resulted in the use of microarray data in application to genetics, medicine, and patient diagnosis. The high dimensional feature vectors in the microarray data often contain large number of features that are not useful in the process of classification. The main role of our feature selection method is to identify the gene expressions from a microarray data that are most useful for classification.

Five benchmark microarray based gene expression databases are used in this study: GLI-85 (also known as GSE4412)[48], GLA-BRA-180 (also known as GDS1962)[49], CLL-SUB-111 (also known as GSE2466)[50], TOX-171 (also known as GDS2261)[51], and SMK-CAN-187 (also known as GSE4115)[52].

Selection threshold and classification

To assess the recognition performance of the proposed feature selection method for the microarray databases listed in Table 1, we randomly select equal number of samples to form the training and test sets. It should be noted that for all the experiments and results presented in this section, a random split of 50% is used for the individual classes in the databases to form the train and test sets. The average recognition accuracies are reported for 30 repeated random splits. The number of features that have an area of overlap within a specified selection threshold can vary from one database to another. This means that the quality of feature can vary in different databases, depending on the level of natural variability within a database. Figure 1 illustrates this observation by the dependencies of the normalized number of selected features z_s on the selection threshold. It can be seen that the quality of the features is different for almost every database. Interestingly, all databases apart from SMK-CAN-187 contain less than 3% of features with a relative overlap area smaller than 0.2. This means that the intra-class variability in SMK-CAN-187 is lower than the other databases, and is possibility because lung cancer affects several gene expressions distinctively in comparisons with other cancer and toxicology databases.

Table 1 Organization of the databases used in the experiments

Database	Number of instances	Number of features	Number of classes	Category
GLI-85 [48]	85	22283	2	Microarray
GLA-BRA-180[49]	180	4915	4	Microarray
CLL-SUB-111[50]	111	11340	3	Microarray
TOX-171[51]	171	5748	4	Microarray
SMK-CAN-187[52]	187	19993	2	Microarray

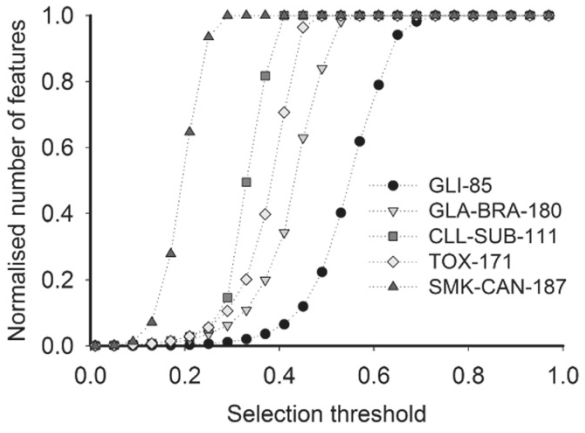

Figure 1 Selection threshold versus selected features. The dependence of selection threshold on the number of selected features for 5 gene expression databases.

Figure 2 shows the recognition performance of the presented feature selection method when used with the nearest neighbor classifier. The recognition accuracy is defined as the ratio between the total number of correctly identified test samples as belonging to a class to the total number of test samples. It can be seen that for all the databases, a selection threshold (σ) of 0.3 or less is sufficient to obtain high recognition accuracies. The maximum values of accuracies are possibly limited by the nature of the classifier and quality of the best features.

Feature ranking and classification

When the relative area of overlap for all the features is small, applying the threshold based selection results in the use of almost all available features for classification. The use of complete set of features in the process of automatic classification is often not a feasible option due to the issues of curse of dimensionality. In such situations, ranking the features and selecting a group of top ranked features can be used for both the dimensionality reduction and selection of the best available features for classification. The simplest and

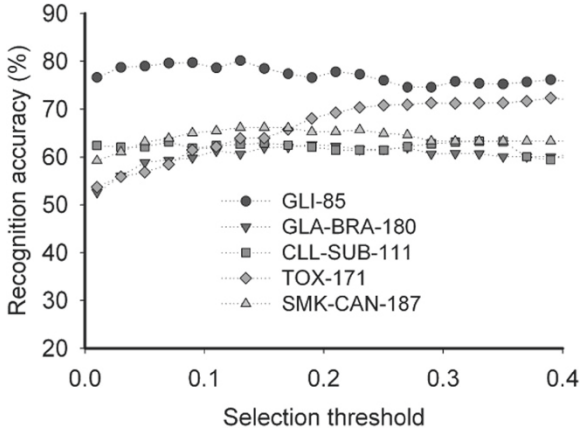

Figure 2 Average recognition performance versus threshold. Average recognition performance of the nearest neighbor classifier used with the newly proposed feature-selection method for 5 gene expression databases.

common approach for selection of the top ranks is by individual searches that evaluate each feature separately. Leave one out cross-validation is performed using the training set of individual features that are selected based on a specified value of selection threshold. The selected features are ranked based on the recognition error by evaluating it individually with a nearest neighbor classifier.

Figure 3 shows the dependence of recognition accuracies on the number of top ranked features used with a nearest-neighbor classifier. This dependence is illustrated for the maximum number of 100 features that all fall below the selection threshold of 0.2 and are ranked based on the least recognition error using the cross validation test. It can seen that a small number of top-ranked features increases the recognition accuracy to the maximum values observed in Figure 2.

Comparisons

Table 2 shows the comparison of the best accuracies obtained with top ranked features using four conventional classifiers: nearest neighbor, linear SVM, and naive Bayes. The recognition accuracies shown in Table 2 is the total number of correctly identified labels of the test samples as belonging to a class in training set to that of the total number of test samples in a test set, where the process of calculating accuracy is repeated for 30 random selections of testing and training set in each of the micro-array databases. Such a cross-validation is done to ensure the correctness of the reported accuracy. The accuracy values of each database is reported on the samples from the testing set using the features selected by the proposed method. Overall, it can be seen that all the classifiers perform equally well. It should be noted here that in most cases, the highest recognition accuracies are obtained with a very small number of features in comparison with the total number of available features. This means that for gene expression databases only very few gene expressions are useful for the process of classification irrespective of the type of classifier employed.

Table 3 shows the pe rformance comparison between the newly presented feature selection method and conventional feature selection methods[53,54]. The accuracy and features are determined using the same process as mentioned for Table 2, It can be seen

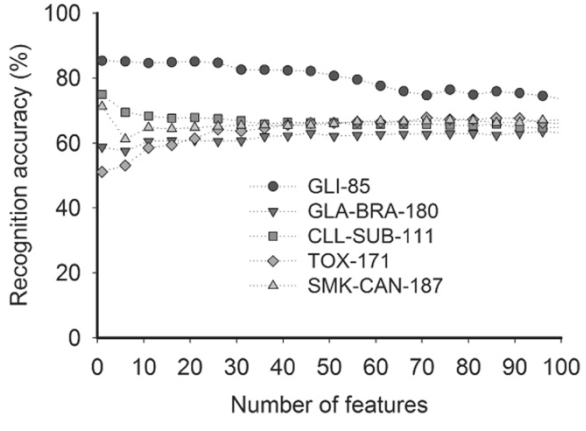

Figure 3 Average recognition performance versus ranked features. Average recognition accuracies obtained by the nearest-neighbor classifier and a selection of up to top 100 features for five different gene expression databases.

Table 2 The highest recognition accuracies on gene expression databases when selecting features within the top 100 ranked features obtained by three different classifiers

		Nearest neighbor		SVM		Naive Bayes	
Database	**Total number of features**	**Accuracy (%)**	**Selected number of features**	**Accuracy (%)**	**Selected number of features**	**Accuracy (%)**	**Selected number of features**
GLI-85	22283	88.3 ± 2.9	3	86.5 ± 5.2	2	**89.1±2.9**	3
GLA-BRA-180	4915	65.3 ± 4.6	45	66.7 ± 4.8	6	**68.4±5.1**	37
CLL-SUB-111	11340	**74.9±2.6**	1	65.6 ± 5.5	78	66.5 ± 8.3	50
TOX-171	5748	69.6 ± 4.4	89	**78.5±5.5**	71	61.5 ± 5.1	68
SMK-CAN-187	19993	71.2 ± 1.7	1	**73.2±3.2**	48	70.8 ± 4.0	52

Table 3 Comparison of maximum recognition accuracies on gene-expression databases using up to 100 top ranked features obtained by different feature-selection methods and a nearest neighbor classifier

		MRMR[53]		Information gain[54]		Presented	
Database	**Total number of features**	**Accuracy (%)**	**Selected number of features**	**Accuracy (%)**	**Selected number of features**	**Accuracy (%)**	**Selected number of features**
CLL-SUB-111	11340	64.5 ± 6.7	32	64.2 ± 8.0	34	**74.9±2.6**	1
SMK-CAN-187	19993	65.1 ± 4.3	41	65.1 ± 3.8	29	**71.2±1.7**	1
GLI-85	22283	83.4 ± 4.5	67	84.2 ± 5.0	87	**88.3±2.9**	3
GLA-BRA-180	4915	64.8 ± 3.4	45	65.6 ± 4.5	27	**68.4±5.1**	37
TOX-171	5748	66.2 ± 5.1	100	65.5 ± 5.0	92	**69.6±4.4**	89

that the presented method uses a fewer number of features to achieve higher recognition accuracies, which shows that the presented method results in more accurate selection of the features that are useful for recognition compared to the conventional methods. The ability of the proposed method to detect fewer number of features without compromising the recognition performance can have a significant impact on the early detection and diagnosis of human diseases (eg glioma) using gene expressions. The detection of such feature imply that they reflect those set of features that indicate the incidence of a particular disease. Any significant change in the such features are indicative of an abnormality or precedence of belonging to a particular state or condition.

Conclusion

In this paper, we presented a feature selection method for gene data classification that is based on the assessment of discriminatory ability of individual features within a class. The area of overlap between inter-class and intra-class distance distributions of individual features is identified as a useful measure for feature selection. A common framework to select the most important set of features is provided by applying a selection threshold. The ability of the proposed method to select the most discriminatory features resulted in improved classification performance with a smaller number of features, although the number of features that are required for achieving high recognition accuracy varies from one database to another. The presented feature selection technique can be used in the automatic identification of cancer causing genes and would help facilitate early detection of specific diseases or conditions.

Competing interests
Both authors declare that they have no competing interests.

Acknowledgements
We would like to thank the anonymous reviewers for their constructive comments which has helped to improve the overall quality of the reported work.

Author details
[1]School of Computer Science, Indian Institute of Information Technology and Management (IIITM) - Trivandrum, Kerala, India. [2]Griffith School of Engineering, Griffith University, Brisbane, Australia.

Received: 20 August 2011 Accepted: 28 May 2012 Published: 24 June 2012

References

1. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Machine Learning Res 3: 1157–1182
2. Saeys Y, Inza I, Larraage P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507–2517
3. Inza I, Larranaga P, Blanco R, Cerrolaza A (2004) Filter versus warpper gene selection approaches in dna microarray domains. Artif Intelligence Med 31: 91–103
4. Ma S, Huang J (2008) Penalized feature selection and classification in bioinformatics. Brief Bioinfrom 9(5): 392–403
5. James AP, Maan A (2011) Improving feature selection algorithms using normalised feature histograms. IET Electron lett 47(8): 490–491
6. Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Boston, Kluwer Academic Publishers
7. Donoho D (2006) Formost large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparest solution. Comm Pure Appl Math 59: 907–934
8. Fan J, Samworth R, Wu Y (2009) Ultrahigh dimensional feature selection: Beyond the linear model. J Machine Learning Res 10: 2013–2038
9. Glocer K, Eads D, Theiler J (2005) Online feature selection for pixel classification. In: 22nd Int Conference Machine Learning. ACM New York, USA, pp 249–256
10. Zhao Z, Liu H (2008) Multi-scource feature selection via geometry dependent covariance analysis. J Machine Learning Res, Workshop Conference Proc Volume 4: New Challenges Feature Sel Data Min Knowledge Discovery 4: 36–47
11. James AP, Dimitrijev S (2012) Nearest Neighbor Classifier Based on Nearest Feature Decisions. Comput J. doi:10.1093/comjnl/bxs001
12. James A, Dimitrijev S (2010) Inter-image outliers and their application to image classification. Pattern Recognit 43(12): 4101–4112
13. Lee JA, Verleysen M (2007) Nonlinear Dimensionality Reduction. New York, Springer

14. Thangavel K, Pethalakshmi A (2009) Dimensionality reduction based on rough set theory: A review. Appl Soft Comput 9(1): 1–12
15. Sanguinetti G (2007) Dimensionality Reduction of Clustered Data Sets. Pattern Anal Machine Intelligence, IEEE Trans 30(3): 535–540
16. Zhao Z, Wang J, Sharma S, Agarwal N, Liu H, Chang Y (2010) An intergrative approach to identifying biologically relevant genes. In: SIAM Int Conference Data Min, pp 838–849
17. Liu H, Yu L (2005) Toward intergrating feature selection algorithms for classification and clustering. IEEE Transactions Knowledge Data Eng 17(3): 1–12
18. Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expressions. Bioinformatics 20(15): 2429–2437
19. Liu H, Li J, Wong L (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genone Inform 13: 51–60
20. Sikonja MR, Kononenko I (2003) Theoritical and emperical analysis of Relief and Relief. Machine Learning 53: 23–69
21. Weston J, Elisseff A, Schoelkopf B, Tipping M (2003) Use of the zero norm with linear models and kernel methods. J Machine Learning Res 3: 1439–1461
22. Song L, Smola A, Gretton A, Brogwardt K, Bedo J (2007) Supervised feature selection via dependence estimation. In: Int Conference Machine Learning. ACM New York, USA, pp 823–830
23. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32: 407–449
24. Zhu J, Rosset S, Hastie T, Tibshirani R (2003) 1-norm support vector machines. In: Adv Neural Inf Process Syst, vol. 16. NIPS foundation, La Jolla, CA p 8
25. Cawley GC, Talbot NLC, Girolami M (2007) Sparse multinomial logistic regression via bayesian L1 regularisation. In: Adv Neural Inf Process Syst, vol. 19. NIPS foundation, La Jolla, CA, pp 209–216
26. Hall MA (2000) Correlation based feature selection for discrete and numeric class machine learning. In: 17th Int Conference Machine Learning. San Fransisco, Morgan Kaufmann, 17:359–366
27. Liu H, Setiono R (1996) A probabilistic approach to feature selection: a filter solution. In: 13th Int Conference Machine Learning, vol. 13. San Fransisco, Morgan Kaufmann, pp 319–327
28. Kohavi R, John G (1997) Wrappers for Feature Subset Selection. Artif Intelligence 97(1-2): 273–324
29. Caruana R, Freitag D (1994) Greedy attribute selection. In: 11th Int Conference Machine Learning, vol. 11. San Fransisco, Morgan Kaufmann, pp 28–36
30. Das S (2001) Filters, warppers and boosting: based hybrid for feature selection. In: 18th Int Conference Machine Learning, vol. 18. San Fransisco, Morgan Kaufmann, pp 74–81
31. Ng AY (1998) On feature selection: learning with exponentially many irrelevant features as training examples. In: 15th Int Conference Machine Learning, vol. 15. San Fransisco, Morgan Kaufmann, pp 404–412
32. Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(3): 131–156
33. Narendra PM, Fukunaga K (1977) Branch and bound algorithm for feature subset selection. IEEE Trans Comput 26(9): 917–922
34. Doak J (1992) An evaluation of feature selection methods and their application to computer security. Tech. rep., University of California, Davis
35. Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Boston, Kluwer Academic
36. Almuallim H, Dietterich TG (1994) Learning boolean concepts in the presence of many irrelavent features. Artif Intelligence 69(1-2): 278–305
37. Ben-Bassat M (1982) Pattern recognition and reduction of dimensionality. Handbook of statistics II, North holand, pp 773-791
38. Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intelligence 97: 245–271
39. Dash M, Liu H (2000) Feature selection for clustering. In: 4th pacific asia conference on knowledge discovery and data mining, pp 110–121
40. Di JG, Brodley CE (2000) Feature subset selection and order itdentification for unsupervised learning. In: 17th Int Conference Machine Learning, vol. 17. San Fransisco, Morgan Kaufmann, pp 247–254
41. Kim Y, Street W, Menczer F (2000) Feature selection for unsupervised learning via evolutionary search. In: 6th ACM SIGKDD international Conference knowledge discovery and data mining, vol. 6. ACM New York, USA, pp 365–369
42. Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 19: 153–158
43. Blum A, Rivest R (1992) Training a 3-Node Neural Networks in NP-Complete. Neural Networks 5: 117–127
44. John GH, Kohavi R, Pflegler K (1994) Irrelavent feature and the subset selection problem. In: 11th Int Conference Machine Learning, vol. 11. San Fransisco, Morgan Kaufmann, pp 121–129
45. Abe S, Thawonmas R, Kobayashi Y (1998) Feature selection by analysing class regions approximated by ellipsoids. IEEE Trans Syst, Man Cybernetics– Part C: App Rev 28: 282–287
46. Neumann J, Schnorr C, Steidl G (2005) Combined SVM-based feature selection and classification. Machine Learning 61: 129–150
47. Xu Z, King I, Lyu MR-T, Jin R (2010) Discriminative semisupervised feature selection via manifold regularization. IEEE Trans. on Neural Networks 21(7): 1033–1047
48. Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, Mischel PS, Nelson SF (2004) Gene expression profiling of gliomas strongly predicts survival. Cancer Res 64(18): 6503–6510
49. Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, James AP (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell 9(4): 287–300
50. Haslinger C, Schweifer N, Stilgenbauer S, Dhner H, Lichter P, Kraut N, Stratowa C, Abseher R (2004) Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. J Clin Oncol 22(19): 3937–3949
51. Piloto S, Schilling T (2010) Ovo1 links Wnt signaling with N-cadherin localization during neural crest migration. Development 137(12): 1981–1990

52. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P, Sridhar S, Beamis J, Lamb C, Anderson T, Gerry N, Keane J, Lenburg ME, Brody JS (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med 13(3): 361–366
53. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Machine Intell 27(8): 1226–1238
54. Cover TM, Thomas JA (1991) Elem Inf Theory. New York, Wiley

doi:10.1186/2192-1962-2-12
Cite this article as: James and Dimitrijev: **Ranked selection of nearest discriminating features.** *Human-centric Computing and Information Sciences* 2012 **2**:12.