Information Processing in Biomedical Applications

Nick J. Pizzi

National Research Council, Institute for Biodiagnostics, Winnipeg MB, R3B 1Y6, Canada
University of Winnipeg, Applied Computer Science Dept., Winnipeg MB, R3B 2E9, Canada

Abstract. To classify biomedical data is to find a mapping from patterns to a set of classes (e.g., disease states). Patterns are represented by features (e.g., metabolite concentrations) and class labels are assigned using a reference test (e.g., an expert's analysis of "normality"). This process often suffers from three significant challenges: voluminous features; pattern paucity; and reference test imprecision. Three computational intelligence based techniques, which exploit the notion of information granulation, are presented to address these challenges. Fuzzy quantile encoding replaces a feature with its membership values in a fuzzy set collection describing the feature's interquantile range. Class label adjustment compensates for reference test imprecision by adjusting design set class labels using a fuzzified similarity measure based on robust measures of class location and dispersion. Stochastic feature selection is a strategy where instances of classifiers are presented with feature regions sampled from an ad hoc cumulative distribution function. These techniques as well as their application to several classification problems in the biomedical domain will be discussed.

Keywords: Biomedical Informatics, Biomedical Data Analysis, Information Granules, Pattern Analysis, Computational Intelligence, Feature Selection, Feature Encoding, Fuzzy Set Theory, Biomedical Data Classification, Artificial Neural Networks, Gold Standards Analysis, Performance Measures, Classifier Aggregation, Fuzzy Integration, Fuzzy Systems, Parallel Computing, Information Processing, Biomedical Applications, Granular Computing, Feature Extraction.

1 Introduction

Human centric computing has as its main objective the development of computing systems that intuitively adjust to the needs of the user in a seamlessly integrated fashion [27]. This paradigm is relevant to a number of information processing fields including pervasive and ubiquitous computing, ambient intelligence, sensor networks, semantic webs, e-health, e-commerce, wearable hardware, and, specific to our case, biomedical informatics. A typical requirement for human centric computing is a "semantic" layer between fine, detailed numerical data and coarser, generalized abstractions. The semantic layer must perform a translation or transformation from data to abstractions in an efficient and effective manner as the data may need to be abstracted in different ways depending on the needs and objectives of a possibly diverse group of users. In the case of biomedical informatics, for instance, it is necessary to provide effective explanatory analysis while simultaneously finding succinct interpretable (biomedically meaningful) representations. Of course, the challenge is determining the underlying semantic translation that provides the optimal mapping from voluminous data to human manageable, qualitative interpretations. One successful approach to deal with this issue is granular computing.

A. Bargiela, W. Pedrycz (Eds.): Human-Centric Information Processing, SCI 182, pp. 289–311.
springerlink.com

Granular computing [26] is an information processing paradigm that deals with complex information entities in a coherent and comprehensive fashion. Central to this theoretical perspective is the concept of information granules – conceptual entities possessing elements of similarity, functional adjacency, or spatial (or temporal) proximity. Information granules are used to describe or interpret phenomena and carry out processing at the level that is most suitable for the designer of the system and most germane to its potential user. In this sense, one may regard granular computing as an important paradigm for the development of human-centric confirmatory (or exploratory) biomedical data analysis. As a rich theoretical perspective, granular computing subsumes and augments the well established disciplines of interval analysis, fuzzy sets, rough sets, and probability theory [24,25,50,51].

Granular computing research focuses on the construction of a coherent conceptual and methodological framework (and related algorithmic issues): granule quantification and discretization; communication mechanisms between environments of different information granularity levels; translation formalisms between granules grounded in different conceptual environments (e.g., possibility–probability transformations or fuzzy/crisp set approximations); granule construction (e.g., via clustering); and analysis/synthesis of granular systems (e.g., granular classifiers).

There are three types of granulation that are germane to biomedical information processing: discretization; conceptual; and clustering. Discretization involves granulation at the level of feature values. This may be achieved by mapping (binning) a range of values for a biological feature (for example, the concentration of a metabolite) to an ordinal value or through rank ordering of feature values. As it is not feasible to examine the effects of all different discretization combinations for a particular biomedical data analysis problem [18], care must be exercised in designing a heuristic to find near-optimal (or at least adequate) discretizations. Concept granulation involves the notion that different sets of features may give rise to interactions leading to different, possibly, conflicting, higher level conceptual formulations. Clustering involves feature aggregation or transformation to reduce the dimensionality of the original biomedical feature space. Many techniques, with varied relative advantages and disadvantages, fall under this category of granulation: multidimensional scaling, agglomerative techniques, principal component analysis, fuzzy clustering, projection pursuit, independent component analysis, factor analysis, and so on [7,26,49]. For biomedical data analysis, it is important that granulation techniques do not mask or diminish the information content present in the original (biomedically relevant) feature space.

1.1 Biomedicine and Vagueness

Imprecision, incompleteness, and uncertainty are intrinsic to the practice of medicine. While this art of making decisions with inadequate information is often impervious to precise modes of analytical reasoning, it is regularly amenable to approximate ones [15], and, as a result, the field has become a fertile and active domain with which to exercise granular computing based modes of reasoning. Further, medical decision-making is paradigmatic of general decision support systems in which principles, procedures, data, and knowledge are approximate; hence, successful methods applied to the medical domain may often be generalized across many application domains.

The medical diagnostic process involves an inference of a disease from a set of symptoms based on a body of medical knowledge about nosology[1] and symptomatology[2]. Unfortunately, vagueness is a hallmark of this process. A disease may manifest itself differently from one patient to the next as well as temporally for the same patient. Medical diagnosis is confounded by multiple diseases present in a particular patient or a specific symptom present in multiple disease states. A patient's historical information may be incomplete, physical examinations may inadvertently ignore symptoms, laboratory test results may be imprecise, and the distinction between the states of normality and abnormality is not necessarily crisp. Important diagnostic information acquired from medical instrumentation such as magnetic resonance, infrared, or mass spectrometers is often complex and voluminous and their interpretation may vary from one expert to the next. Medical knowledge is often couched in necessarily imprecise linguistic terminology. The proliferation of new medical knowledge, introduces uncertainty and inconsistencies during its assimilation into the current orthodoxy. For instance, the inclusion of new diagnostic procedures after they have been successfully assessed against the corresponding external reference test (this currently accepted diagnostic procedure is often referred to as the "gold standard"), which itself may be imprecise.

1.2 Granular Computing and Computational Intelligence Strategy

Classifying biomedical data involves finding a mapping (relationship) from patterns (e.g., data relating to some type of tissue or biofluid) to a set of classes (e.g., disease states). Patterns are represented by features (e.g., concentrations of biological compounds) and class labels are assigned using a reference test (e.g., a medical expert's analysis of tissue being "normal" or "abnormal"). This process often suffers from three significant challenges: the number of features in a pattern is high; the number of patterns is low; and the reference test may be unreliable. The first two challenges, known collectively as the "curse of dimensionality", cause an inability to find robust, general solutions. This is often addressed by reducing, in some fashion, the number of features; however, a direct correspondence back to the original features is necessary for medical experts to make informed judgments about the mapping's predictive power. While external reference tests may be well-established benchmarks, they are seldom perfectly accurate and sometimes improperly applied. Nevertheless, any strategy that compensates for this reference test imprecision must ensure that the mapping is correctly validated against the benchmark. In this chapter we present three techniques based upon computational intelligence and the paradigm of granular computing to deal with the biomedical data analysis challenges described above.

Fuzzy quantile encoding is a classification preprocessing method that replaces a specific feature value for a pattern with its membership values in a collection of fuzzy sets describing the interquantile range of all values for that feature within the dataset. This method "normalizes" features to the unit interval, diminishes the impact of feature "outliers", and improves the overall accuracy and computational performance of adaptive classifiers such as supervised feed-forward neural networks.

[1] The branch of medical science dealing with the classification of diseases.

[2] The study of symptoms of disease and signs of pathogens for the purpose of diagnosis.

Gold standard class label adjustment is a set of mitigation strategies that compensates for possible reference test imprecision by adjusting the design set class labels using a fuzzified similarity measure based on robust measures of location and dispersion of class medoids (robust centroids). These mitigation strategies fall into three categories: reassignment involves changing the class label of a design subset pattern, if it is found to be more "similar" to patterns from another class; surrogation involves using a new space of class labels for the design set (for instance, cluster analysis may indicate that patterns in a particular class are distributed in such a way that they are better represented by two surrogate class labels); and gradation involves the fuzzy set notion of a pattern belonging to all classes to varying degrees (that is, moving from a crisp, Boolean class assignment to a fuzzy one).

Stochastic feature selection is a parallelized classification strategy where many instances of heterogeneous classifiers are presented with (possibly quadratically transformed) feature regions of varying cardinality. Regions are stochastically sampled from an ad hoc cumulative distribution function that is iteratively updated based on a frequency histogram of features used by prior classifiers whose performance (accuracy) exceeds a pre-defined threshold. Fuzzy integration is used to aggregate the best classification outcomes.

The schema presented in Figure 1 indicates the three main classification phases: pre-processing, fuzzy quantile encoding and gold standard class label adjustment;

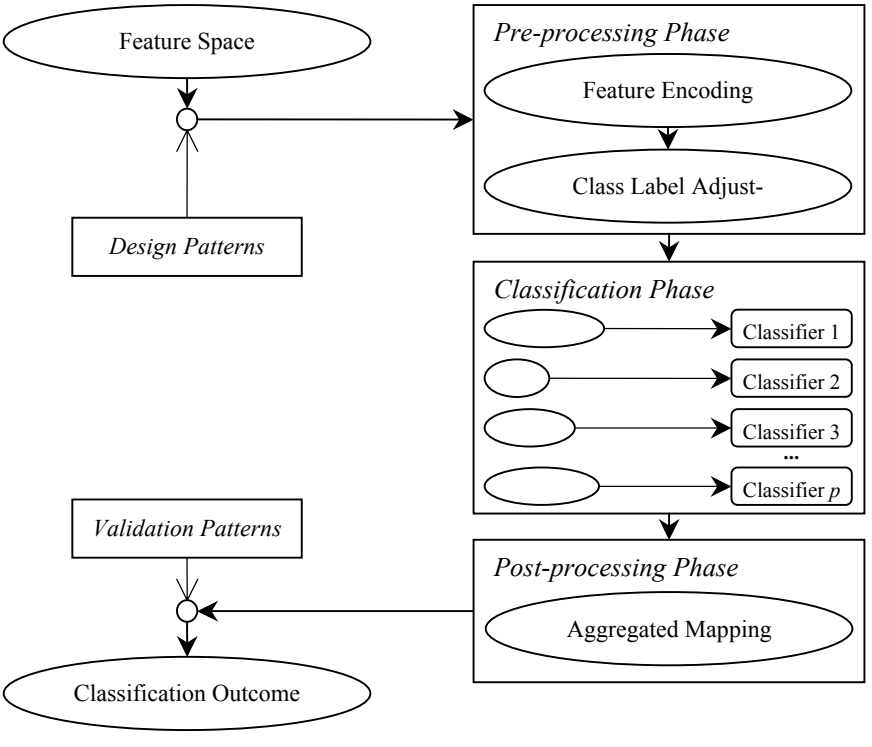

Fig. 1. Schema indicating the pre-/post-processing and classification phases

classification, stochastic feature selection coupled with a set of p heterogeneous classifiers; and, post-processing, mapping aggregation (prediction fusion). The figure also clearly indicates that only design patterns (those patterns randomly assigned to the design subset) are used in these phases to construct the aggregated classification mapping. In order to attenuate bias, classification performance (accuracy) is assessed using this mapping with only the validation patterns.

Each of these classification techniques and strategies will be discussed in the sections that follow. A description of the architecture of the biomedical data analysis software, which implements several key aspects of this methodology, will be provided. Finally, the successful application of this methodology to several classification problems in the biomedical domain will be discussed.

2 Biomedical Data Classification

The latest biomedical spectroscopic modalities produce information rich but complex and voluminous data [23]. For instance, magnetic resonance (MR) spectroscopy, which exploits the interaction between an external homogenous magnetic field and a nucleus that possesses spin, is a reliable and versatile spectroscopic modality [10]. Coupled with robust multivariate discrimination methods, it is especially useful in the classification and interpretation of high-dimensional biomedical spectra of biofluids and tissues [42]. However, the sample to feature ratio of these data is typically low; the feature space dimensionality is $O(10^3–10^4)$ while the sample size is O(10–100). This "curse of dimensionality" is a serious challenge for the classification of biomedical spectra: the excess degrees of freedom tend to cause overfitting, which affects the reliability of the chosen classifier.

Advances in pattern recognition, computational intelligence, and granular computing, contribute ever more sophisticated models upon which to build ever more sophisticated classifiers. Herein lies a major problem: if these models are highly non-linear, they may be unstable, if they are iterative, they may not converge, if they are probabilistic, they may be based on underlying statistical assumptions that are often not true in real-world scenarios. Preprocessing may address these concerns: data may be transformed such that a non-linear model may be replaced by a linear one, the dimensionality of the data may be reduced so that an iterative method may converge or may be substituted for an analytic one, or the data may be "normalized", in some sense, such that the underlying statistical assumptions of a probabilistic model are realized. Years of investigations in biomedical data analysis have led to this author's conjecture that the 80/20 rule holds in the development of classification systems: 20% of a researcher's effort should be spent on selecting and tuning a classifier; 80% should be spent on a thorough analysis of the data to simplify them, via pre-processing, prior to presentation to the classifier of choice.

Here, we formally introduce some notation used throughout this chapter. Let $\boldsymbol{X}=\{(\boldsymbol{x}_k,\omega_k),\ k=1..N\}$ be a set of N patterns, $\boldsymbol{x}_k\in\Re^n$, with respective class labels, $\omega_k\in\Omega$, $\Omega=\{1..c\}$ that are randomly assigned to either a design subset, $\boldsymbol{X}^D$, comprising N_D patterns, or a validation subset, $\boldsymbol{X}^V$, comprising N_V patterns ($N_D+N_V=N$). Classification involves finding a mapping (function approximation), $f:\boldsymbol{X}^D\rightarrow\Omega$, and then validating its effectiveness using $\boldsymbol{X}^V$, $f:\boldsymbol{X}^V\rightarrow\Omega$ (if the predicted class label does not match the assigned class label then it is considered to be a misclassification).

Many classification architectures exist (granular classifiers, supervised artificial neural networks, multivariate statistical methods, evolutionary computation, hybrid strategies, and so on), with various advantages and disadvantages [7]. However, as mentioned above significant effort must be expended in the analysis of appropriate pre-processing strategies. For instance, feature selection is a typical preprocessing strategy for attenuating the effects of the curse of dimensionality by reducing the size of the input (feature) space. Feature selection involves finding a mapping, f:X→X′, where $X' \subseteq \Re^m$ ($m<<n$) is the reduced feature space. Subsequently, classification involves finding a mapping from the reduced feature space to the space of class labels, g: X′→Ω. The intent of this strategy is to select those features possessing significant discriminatory power.

One aspect of biomedical data classification that is often glossed over is the reliable validation of the accuracy results generated by a classification schema. It is essential that datasets be divided (randomly) into design and validation sets. Design patterns may be used in the construction of a classification system but once this phase is complete, performance results must be based on the validation patterns. Given this necessary condition, how is the performance of a classification system to be measured given a $c \times c$ confusion matrix of desired versus actual class labels using the validation patterns? The conventional performance measure is the ratio of correctly classified patterns to the total number of patterns, P_O

$$P_o = N_V^{-1} \sum\nolimits_i n_{ii} \, (i = 1, \ldots, c) \tag{1}$$

where n_{ij} is the number of class i validation patterns predicted to belong to class j. An alternate performance measure is the average class-wise accuracy, P_A

$$P_A = c^{-1} \sum\nolimits_i \left(n_{ii} \Big/ \sum\nolimits_j n_{ij} \right) (i, j = 1, \ldots, c) \tag{2}$$

But neither P_o nor P_A take into account any agreement due to chance [8], P_L

$$P_L = N^{-2} \sum\nolimits_i \left(\sum\nolimits_j n_{ij} \sum\nolimits_j n_{ji} \right) (i, j = 1, \ldots, c) \tag{3}$$

A more conservative performance measure is the κ score [9], a chance-corrected measure of agreement between the desired and actual class assignments

$$\kappa = \left(P_o - P_L \right) \Big/ \left(1 - P_L \right) \tag{4}$$

If the agreement is due strictly to chance, κ=0. If it is greater than chance κ>0; κ=1 indicates complete agreement. If the agreement is less than chance then κ<0 (floor depends upon the marginal distributions). A useful benchmark for agreement strength (confidence) is: poor (κ=0), slight (0.0<κ≤0.2), fair (0.2<κ≤0.4), moderate (0.4<κ≤0.6), substantial (0.6<κ≤0.8), and almost perfect (0.8<κ<1.0) [17]. Table 1 shows the necessity of careful analysis of accuracy. It lists two confusion matrices with the same number of patterns per class and the same overall accuracy, P_O=0.66. Using P_A it is clearer that the accuracy is in fact worse in Table 1(ii), P_A=0.33 than Table 1(i) P_A=0.62. However, via κ, it is clear that the apparent accuracy in the second

Table 1. Two three-class confusion matrices with the same pattern distributions

	(i)C1	C2	C3	(ii)C1	C2	C3	N=300
C1	15	10	5	3	24	3	N_1=30
C2	37	163	40	24	192	24	N_2=240
C3	2	8	20	3	24	3	N_3=30
Accuracy	P_O=0.66	P_A=0.62	κ=0.29	P_O=0.66	P_A=0.33	κ=0.00	

confusion matrix is due strictly to chance, κ=0.00 versus κ=0.29. This is further evidenced by examining P_L (0.52 versus 0.66).

Here, we briefly present three classifiers (two neural networks and one statistical method) that have been used for biomedical data analysis described in the applications section. Neural networks [1] are self-adaptive, machine learning systems composed of layers of processing elements, which are sets of inputs and weights combined to generate outputs used by an adjacent layer. Supervised networks [38] require the desired class labels for each pattern so that they may be compared to the predicted label. Based on these comparisons, a learning strategy, used to make incremental changes to the weights, minimizes an error criterion.

The multi-layer perceptron (MLP) [39] is a supervised feed-forward network, which has consistently demonstrated its effectiveness as a reliable nonlinear classification technique [3]. The transfer function γ (often the logistic function, $\gamma(x)=(1+e^{-x})^{-1}$) is sigmoidal and the output of processing element j is $x_j=\gamma(\sum_i w_{ji}x_i)$. In general, an MLP may be considered a non-linear regression system that performs a gradient descent search through the weight space, searching for minima.

The probabilistic neural network (PNN) [43] uses patterns to construct probability density functions (pdf) to estimate the likelihood of a given pattern belonging to a class. When the class pdfs are known, a PNN correspond to a Bayesian classifier. Since true class pdfs are rarely known, they are usually approximated via a sampling histogram and Parzen estimators [22]. This involves the construction of unit area Gaussians centred at the values of the features for every design pattern. These Gaussians are summed and scaled to produce a composite curve. As the number of design patterns increase, the composite curve asymptotically approaches the true pdf. [However, it is not possible to determine the number of patterns required to estimate the pdf to a specified accuracy.]

Linear discriminant analysis (LDA) is a classification approach that determines linear decision boundaries between c classes while taking into account inter- and intra-class variances [40]. If the error distributions for each class are the same, LDA constructs the optimal linear decision boundary between the classes. In real-world situations, this optimality is seldom achieved since different classes typically give rise to different distributions. LDA is a useful linear classifier; however, when appropriate data preprocessing is applied, in particular, dimensionality reduction techniques such as stochastic feature selection. LDA allocates a pattern, $\boldsymbol{x}$, to class i for which the probability distribution, $p_i(\boldsymbol{x})$, is greatest. That is, $\boldsymbol{x}$ is allocated to class i, if $q_i p_i(\boldsymbol{x})>q_j p_j(\boldsymbol{x})$ $(\forall j\neq i)$, where q are the prior (or proportional) probabilities. The discriminant function is $L_i(\boldsymbol{x})=\log q_i+\boldsymbol{m}_i^{\mathrm{T}}W^{-1}(\boldsymbol{x}-\frac{1}{2}\boldsymbol{m}_i)$ where $\boldsymbol{m}_i$ is the mean for class i and $\boldsymbol{W}$ is the covariance matrix.

3 Stochastic Feature Selection

Stochastic feature selection (SFS) is a feature selection/reduction pre-processing method that is tightly coupled to the classification phase. SFS may be used with any homogeneous or heterogeneous set of classifiers (e.g., LDA, MLP, PNN, or support vector machines [47]). Essentially, SFS iteratively presents, in a highly parallelized fashion, many feature regions (contiguous subsets of pattern features) to the set of classifiers retaining the best set of classifier/region pairs. Figure 2 lists several of the key parameters used in SFS, which we will reference in the following detailed description of SFS.

After selecting the minimum and maximum number of feature regions and the minimum and maximum size (cardinality) for a feature region (cf. fields shown in Figure 2, "Min number of regions", "Max number of regions", "Min region length", "Max region length", respectively), the general procedure is: (i) randomly select a number of feature regions and, for each region, select a random size (satisfying the above constraints); (ii) prune the features not selected in (i) from the training and monitoring sets; (iii) use the training set and classifier to produce classification coefficients; (iv) test these candidate coefficients with the monitoring set; (v) repeat steps (i)–(iv) until either the accuracy threshold ("Fitness threshold") or maximum number of iterations ("Max number of iterations") is exceeded; (vi) finally, use the best coefficients found and assess their performance using the validation set. Note that the training and monitoring sets are composed of patterns exclusively from the design set. The validation patterns are only used in step (vi).

SFS retains a list of the best classification results ("Number of results to return/keep") based on the selected fitness function ("Order by"). The fitness function

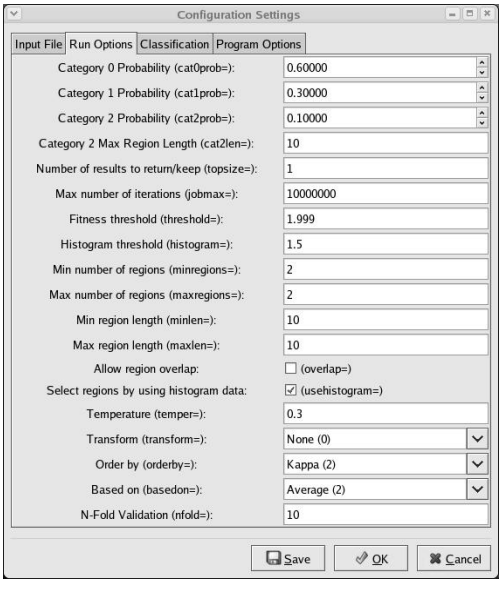

Fig. 2. Several parameters used for stochastic feature selection

may be P_O, P_A, or κ, which may be applied ("Based on") exclusively to the training set or in conjunction with the monitoring set and internal cross-validation is also used ("N-Fold Validation"). Feature regions are normally disjoint but this can be relaxed ("Allow region overlap"). Moreover, transformations may be performed on regions ("Transform") such as computing their average feature value, their variance, or other statistical moments.

3.1 Feature Frequency Histogram

The stochastic nature of this method is normally controlled by the feature frequency histogram (see Figure 3). During an SFS run, the performance of each classification task is assessed using the selected fitness function. If the fitness exceeds the histogram fitness threshold (cf. Figure 2, "Histogram threshold"), which is set to some value less than the fitness threshold stopping criterion, the frequency histogram is incremented at those feature indices corresponding to the regions used by the particular classification task. This histogram is then used to generate a cumulative distribution function (cdf). Now, when feature regions are selected for a new classification task, features are randomly selected using the current cdf. So, rather than each feature having an equal likelihood of being selected for a new classification task, those features that were used in previous "successful" tasks have a greater likelihood of being chosen. A temperature term, $t \in [0,1]$, provides additional control over this process. If $t=0$, the cdf is used as described but, as $t \to 1$, the randomness becomes more uniform (when $t=1$ a strict uniform distribution is used). A useful interactive option is to pause SFS, select those regions that have been shown to be most discriminatory, and continue SFS so that subsequent regions will be selected only from these highly discriminatory features.

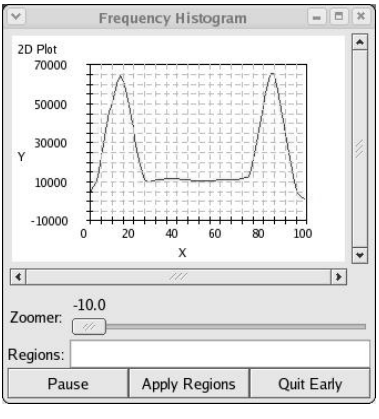

Fig. 3. A typical SFS feature frequency histogram

3.2 Quadratic Combination of Features

A useful SFS pre-processing option is to augment the original features with a quadratic combination of feature regions. The intention here is that if the original feature

space possesses non-linear decision boundaries between classes, the new (quadratic) parameter space may possess more "linearized" decision boundaries. For instance, say we have a set of three-feature two-class points (patterns), $\boldsymbol{x}=\{x_1,x_2,x_3\}\in[0,1]^3$) bounded by the unit hypercube where one class of points, ω_1, are those within the unit hypersphere ($x_1^2+x_2^2+x_3^2<1$) and the other class, ω_2, are those points outside ($x_1^2+x_2^2+x_3^2\geq 1$). These patterns are obviously separated by a circular (non-linear) decision boundary. A linear classification system using, for instance, linear discriminant analysis, would perform poorly ($P_O\approx 0.50$) with such a dataset as no linear decision boundary (plane) can accurately delineate the two classes of points (patterns). However, if we create a new three-coordinate feature space by simply squaring the original features, the decision boundary (in this new space) would be a plane and a linear classifier will now perfectly separate the two classes of patterns.

SFS has three categories of quadratic combinations with which to augment the original features (cf. the respective fields shown in Figure 2): (i) using the original feature region ("Category 0 Probability"); (ii) squaring the values for the selected feature region ("Category 1 Probability"); or (iii) using all pair-wise cross-products of features from two regions ("Category 2 Probability"). Given the potential combinatorial explosion with the third category, an upper limit for the region size may also be specified ("Category 2 Max Region Length"). The probabilities of selecting one of these quadratic combination categories must sum to 1.0.

3.3 Parallelized Classification

SFS takes full advantage of parallel computations using the Scopira Agent Library [6], a sophisticated message-passing library similar in functionality to MPI [41]. Given a high-performance computing cluster (e.g., a Linux Beowulf cluster) environment, classification tasks are distributed to slave nodes for computation. A master node coordinates the distribution of tasks, updates the feature frequency histogram, and records intermediate classification performance results. To minimize interprocess communication and maximize continuous computational loads on the processors, SFS efficiently "bundles" sets of classification tasks. Furthermore, while SFS exploits parallelism, it still remains a strictly deterministic system. That is, experimental results are perfectly reproducible regardless of computational load, which is extremely important in the analysis, and interpretation of complex biomedical data.

4 Fuzzy Quantile Encoding

Zadeh's seminal work on fuzzy set theory [51] may be applied to a classification preprocessing technique that encodes the feature space prior to presentation to a classifier. For instance, a feature may be intervalized across a collection of fuzzy sets thereby producing a list of degrees of membership for each of the fuzzy sets. In other words, given s fuzzy sets, F_1, F_2, …, F_s, and f_i is the membership function for fuzzy set i, then the list of values for a single feature value x is $\{f_1(x), f_2(x), \ldots, f_s(x)\}$. Figure 4 illustrates this intervalization approach using the membership functions for two fuzzy sets for feature i that overlap at 0.5 (see below).

Fuzzy quantile encoding (FQE) uses a feature's quantile values as the consecutive intersections of triangular (or trapezoidal) fuzzy sets [36]. To derive the formula (a full derivation and complete discussion may be found in [30]), the following terms need to be defined. Let b, $0 \le b \le 1$, be the boundary value at the intersection of the fuzzy sets. For simplicity, b may be held constant for each intersection. Let w be the width of the top of the trapezoid of the fuzzy sets. [If w=0, the f_i's are triangular fuzzy sets.] Let l_i and r_i be the left and right boundary, respectively, of the fuzzy set F_i such that $f_i(l_i)=f_i(r_i)=b$. Finally, let x be the original non-encoded input value. Then,

$$f_i(x) = \begin{cases} 1 \wedge \left(0 \vee \left(1 + w - 2\dfrac{1+w-b}{r_i - l_i} \left| x - \dfrac{l_i + r_i}{2} \right| \right) \right) & l_i < r_i \\ 1 & l_i = r_i = x \\ 0 & l_i = r_i \neq x \end{cases} \tag{5}$$

where $\vee$ and $\wedge$ are the max and min operators, respectively (other norm/co-norm pairs, of course, are permissible). The latter two cases define a delta function when $l_i=r_i$. These delta functions satisfy the criteria for a fuzzy set: it is monotonic and it maps onto the unit interval. Delta functions may arise when pattern feature values are significantly skewed (non-normal). It is important to note that, since $f_i(r_i)=f_{i+1}(l_{i+1})=b$, $r_i=l_{i+1}$ ($\forall i$=1..s–1). It should also be noted that the corresponding membership functions are symmetric about the boundaries l_i and r_i. When $b \ge 0.5$ and w=0 there exists a strict 1–1 correspondence between the encoding and the original feature value. When b<0.5 (or w>0), a 1–many correspondence exists.

Quantiles are used to determine reasonable values for the fuzzy set boundaries l_i and r_i. The Q^{th} quantile of N feature values is a value such that Q% of the area under the relative frequency distribution for the feature values lies to the left of the Q^{th} quantile and (100–Q)% of the area under the distribution lies to its right.

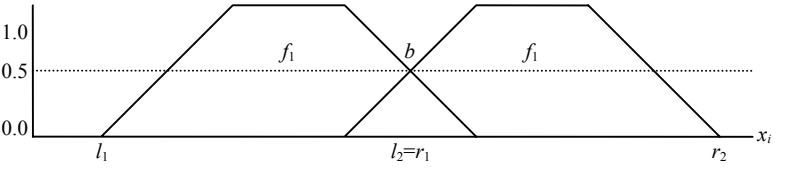

Fig. 4. FQE membership functions using two fuzzy sets for feature i

4.1 Interquartile Range

Normally, the selected quantiles for FQE are the feature's quartiles (see Figure 5 illustration): the lower quartile (25$^{\text{th}}$ quantile), Q_L; the median (50$^{\text{th}}$ quantile), m; and the upper quartile (75$^{\text{th}}$ quantile), Q_U. By using the interquartile range for the feature j,

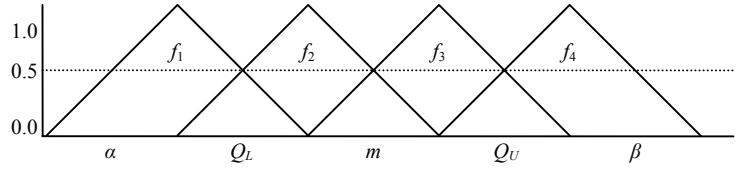

Fig. 5. FQE membership functions using a feature's interquartile range

uniform coverage is effected through the use of four overlapping fuzzy sets, F^j_1, F^j_2, F^j_3, F^j_4. To ensure a 1–1 mapping between the original feature values and the FQE values, w=0 and b=0. [However, the constraint on w can be relaxed (see sub-section below).] Specifically, the membership functions for feature j are

$$\begin{aligned}
f_1^j(x_j) &= 1 \wedge \left[0 \vee \left[1 - \left|x - 0.5\left(\alpha^j + Q_L^j\right)\right| \Big/ \left(Q_L^j - \alpha^j\right)\right]\right] \\
f_2^j(x_j) &= 1 \wedge \left[0 \vee \left[1 - \left|x - 0.5\left(Q_L^j + m^j\right)\right| \Big/ \left(m - Q_L^j\right)\right]\right] \\
f_3^j(x_j) &= 1 \wedge \left[0 \vee \left[1 - \left|x - 0.5\left(m^j + Q_U^j\right)\right| \Big/ \left(Q_U^j - m^j\right)\right]\right] \\
f_4^j(x_j) &= 1 \wedge \left[0 \vee \left[1 - \left|x - 0.5\left(Q_U^j + \beta^j\right)\right| \Big/ \left(\beta^j - Q_U^j\right)\right]\right]
\end{aligned} \tag{6}$$

where α^j and β^j are the feature's respective minimum and maximum values. A dimension-preserving variant to this fuzzy encoding approach [29] is to use a single membership function, $f_j(x)$, which corresponds to a piece-wise linear fuzzy set (w=0), to capture the information represented by the feature's interquartile range

$$f^j(x) = \begin{cases}
b(x-\alpha)(Q_L-\alpha)^{-1} & \text{if} \quad \alpha \le x < Q_L \\
(1-b)(x-Q_L)(m-Q_L)^{-1} + b & \text{if} \quad Q_L \le x < m \\
(b-1)(x-m)(Q_U-m)^{-1} + 1 & \text{if} \quad m \le x < Q_U \\
-b(x-Q_U)(\beta-Q_U)^{-1} + b & \text{if} \quad Q_U \le x < \beta \\
0 & \text{if} \quad x < \alpha \vee x > \beta
\end{cases} \tag{7}$$

4.2 Dispersion Adjustment

An effective extension to fuzzy quantile encoding involves adjusting the fuzzy sets in order to take into account a feature's overall dispersion of values [28]. A robust technique to implement dispersion-adjusted FQE is to use a feature's median of absolute deviations, τ

$$\tau(x) = \frac{m(|x - m(x)|)}{\sigma} \tag{8}$$

where m is the feature's median and σ=0.6745 to ensure that, as the error distribution becomes more normal, τ converges to the standard deviation. While only 40% efficient

for normal data [13], τ is robust to outliers and long-tailed distributions. In other words, as the features becomes more contaminated (less normal), the relative efficiency of τ becomes greater than the standard deviation.

In order to take into account a feature's overall dispersion, the constraint on w needs to be relaxed; for a given pattern feature, let $w=\tau$. Using (8), (5) can easily be modified to now permit the use of trapezoidal fuzzy sets. As the dispersion increases (τ becomes larger), the width of the trapezoid increases and, as a result, more original feature values will be encoded to 1. As the dispersion decreases, the trapezoid approaches a triangular fuzzy set, so fewer values will be encoded to 1.

4.3 FQE Properties

FQE may be easily integrated into any classification system. The input layer (feature space) will have $s \times n$ coordinates where n is the dimensionality of the original feature space and s is the number of fuzzy sets used for encoding (s=4 for interquartile encoding). FQE exhibits several useful properties.

First, the feature space is "normalized": that is, for any given pattern feature, x, its corresponding membership functions map feature values onto the unit interval, $f_i(x) \in [0,1]$ $(\forall i=1..s)$. This is particularly useful during the classification process since scaled biomedical data stabilize the effects of extreme variance disparities across pattern features [38]. Without scaled data, features with large variances will have a tendency to predominate, during the training phase, over those features with small variances even though the latter features may be highly discriminatory.

Another beneficial property is that, during the construction of the discriminating class decision boundaries, feature values that may be considered as outliers impact less severely upon classifiers that employ any type of iterative adjustments to its error function (e.g., artificial neural networks such as MLP). This does not mean that patterns with features that are outliers are removed during the design or validation phases of the classification process, however. FQE values will approach zero as values move outside a feature's interquartile range. In the case of MLP, where its hidden layer processing elements are summing products of weights and input values this is important since, if the FQE values of an outlier are all zero or near zero, those values will contribute very little to the learning process (local error adjustments) regardless of the processing elements weights. This is an extremely useful property if the original feature value is indeed an outlier (nevertheless, if it is not an outlier it still does contribute to a degree). Conversely, values that are within the feature's interquartile range will contribute strongly to the iterative learning process.

Another purpose behind FQE intervalization, as with any type of intervalization, is to reduce the effects of noise in the data as well as to transform the problem in such a way that a non-linear regression model such as MLP can provide better (more accurate) solutions.

Moreover, a FQE based classifier projects the original n-dimensional pattern feature space onto a new $4n$-dimensional parameter space of membership values. This projection often has the positive effect of "linearizing", to some degree, the discrimination problem (that is, moving from non-linear to linear class decision boundaries).

Further, since many FQE values are zero (or near zero), artificial neural network processing elements that use these encoded values as input terms will produce output values that are also at or near zero regardless of the corresponding processing element weights. Subsequently, these processing elements tend to contribute little to the overall classification error (and, derivatively, to the overall learning) of the FQE-based artificial neural network so resultant errors propagated back through the neural network are not caused (to any great extent) by these values. These simplifications, caused by the projection, often significantly reduce the training phase convergence time for supervised artificial neural networks [29].

5 Fuzzy Class Label Adjustment

Gold standard fuzzy class label adjustment (GSA) compensates for the possible imprecision of a well-established but tarnished gold standard (external reference test) by adjusting, if necessary, the class labels of the design set patterns. The procedure begins with finding the centroids of each class using their respective design set patterns. Distances are computed between each design pattern and each class centroid. A fuzzy set theoretic membership function uses these distances to adjust the class labels; in general, the further a pattern is from a class centroid, the lower its membership value for that class. However, the class label for a pattern will only be adjusted if it is sufficiently distant from the centroid of its original class and sufficiently near another class' centroid. Note that any adjustments made to the gold standard occur only for patterns in the design set; for verification purposes, the class labels for the validation set patterns are never altered. Hence, the efficacy of this method is always measured against the original gold standard (regardless of its possible imprecision).

Distances and dispersions are measured using robust multivariate statistics since they are much more resistant to effects caused by extreme feature values than parametric statistics. More specifically, a statistical estimate is robust if it is insensitive to slight deviations from its requisite model assumptions (often normal assumptions) about the underlying feature distribution [13]. This is crucial when dealing with outliers, patterns that do not follow the distribution of the majority of the data.

Although it is a univariate estimator, $\tau(x)$ (see (8)) may be extended to the multivariate case by computing a vector, $\boldsymbol{\tau}^l=[\tau^l_1..\tau^l_n]$, which is a feature-wise measure of dispersion for the class l patterns. First, let $\boldsymbol{X}^l=\{(\boldsymbol{x}_k,l),\ k=1..N_l\}\subset\boldsymbol{X}$ be the set of all patterns belonging to class l (N_l is the number of class l patterns and $\boldsymbol{x}_k=[x_{k1}..x_{kn}]$). Also, let $\boldsymbol{z}^l_j=[x_{ij}]$ ($i=1..N_l$) be the respective values of feature j for the class l patterns. Now, $\tau^l_j=m|\boldsymbol{z}^l_j-m(\boldsymbol{z}^l_j)|/0.6745$. The distance between each pattern and each of the class centroids may now be determined. The weighted distance, d^l, of $\boldsymbol{x}_k$ from the class l centroid (more correctly its medoid) may be defined as

$$d^l(\boldsymbol{x}_k)=\sum_j\left|\frac{x_{kj}-m(z^l_j)}{\tau^l_j}\right| \tag{9}$$

This distance measure is incorporated into the original gold standard using membership functions; the class l membership function for a pattern, $\boldsymbol{x}_k$, is defined as

$$f_l(x_k) = \left[1 + \left(d^l(\boldsymbol{x}_k)/q\right)^p\right]^{-1} \tag{10}$$

where $p>1$ and $q>0$ describe the shape and amount of fuzziness for the membership function ($0 \le f_l(\boldsymbol{x}_k) \le 1$). Figure 6(i) plots (10) for different values of p with a constant q. Note that f is sigmoidal and that as p increases, f approaches a step function. The point, at which the membership function is 0.5, occurs when the distance equals q. Figure 6(ii) plots f for different values of q with a constant p. As q increases, membership values will remain high even at great distances.

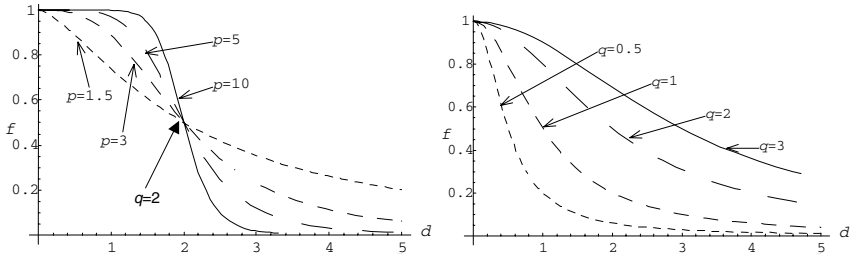

Fig. 6. Plot of f versus distance (d) between a pattern and a class medoid with (i) varying p (q=2) and (ii) varying q (p=2)

Finally, we use contrast intensification, y_l, on the class l membership function to increase membership values above 0.5 and reduce those values that are below this point [51]. Using GSA, we may now recode the class label for $\boldsymbol{x}_k$ from the scalar ω_k to the vector $[y_i]$ (i=1..c).

$$y_l(\boldsymbol{x}_k) = \begin{cases} 2f_l^2(\boldsymbol{x}_k) & \text{if } 0 \le f_l(\boldsymbol{x}_k) \le 0.5 \\ 1 - 2\left(1 - f_l(\boldsymbol{x}_k)\right)^2 & \text{if } 0.5 \le f_l(\boldsymbol{x}_k) \le 1.0 \end{cases} \tag{11}$$

If $\boldsymbol{x}_k$ was originally assigned to, say class l, by the gold standard, it may be the case that it was, in fact, closest to some other class medoid, say class o. In this case, $y_o(\boldsymbol{x}_k) > y_l(\boldsymbol{x}_k)$ and, hence, the original gold standard assignment will no longer predominate. If this is undesirable (or unacceptable) for the particular problem domain, the situation may be rectified by constraining the membership functions expressed by (11) so that $f_l(\boldsymbol{x}_k) = f_o(\boldsymbol{x}_k) + \varepsilon$ where ε is a small positive constant. Now, a pattern will never be reassigned to a class different from the one to which it was originally assigned. However, if a pattern is sufficiently near another class medoid then the corresponding class membership value for that pattern will not be zero. In general, the further $\boldsymbol{x}_k$ is from a class medoid, the lower its membership value for that class. While the original class label assigned by the gold standard is crisp ($\boldsymbol{x}_k$ belongs to one and only one class with degree 1), the (soft) class label assigned by GSA (using (11)) is fuzzy ($\boldsymbol{x}$ belongs to all classes to varying degrees).

6 Classifier Aggregation

The fuzzy measure [44] is a set function used to express the grade of fuzziness. Say, $\boldsymbol{X}$ is a non-empty universe of discourse and B, is a σ-field of $\boldsymbol{X}$ [14]. Given the sets, A_1 and A_2, B is a family of subsets of $\boldsymbol{X}$ if: (i) $\varnothing \in B$; (ii) $\boldsymbol{X} \in B$; (iii) if $A_1 \in B$ then $\neg A_1 \in B$; and, (iv) B is closed under set union (i.e., if $A_1 \in B$ and $A_2 \in B$ then $A_1 \cup A_2 \in B$)]. The set function, $g{:}B \rightarrow [0,1]$, is a fuzzy measure over $\boldsymbol{X}$ if three axioms hold: (i) $g(\varnothing)=0$ and $g(\boldsymbol{X})=1$ (boundary conditions ensure that regardless of the degree of evidence an element must not belong to the null set and it must belong to the universe of discourse); (ii) if $A_1, A_2 \in B$ and $A_1 \subset A_2$ then $g(A_1) \leq g(A_2)$ (evidence of an element's membership in a set must always be at least as great as that in any of the set's subsets); and, (iii) if $A_1 \in B$ and A_1 is monotone increasing then $\lim g(A_1) = g(\lim A_1)$ (consistency constraint). A fuzzy measure commonly found in the literature is the Sugeno fuzzy measure [45], g_λ, which satisfies the additional constraint that $g(A_1 \cup A_2) = g(A_1) + g(A_2) + \lambda g(A_1) g(A_2)$, where $\lambda > -1$ and $A_1 \cap A_2 = \varnothing$.

The fuzzy integral [11] is a nonlinear aggregation scheme for combining multiple sources of information to arrive at a "confidence value" for a decision (hypothesis). Let us define a mapping $h{:}\boldsymbol{X} \rightarrow [0,1]$ where a finite ordered $\boldsymbol{X}=\{x_1 \ldots x_n\}$ is of interest. Typical examples are the Sugeno, $Su(x)$, Choquet, $Ch(x)$, and Shilkret, $Sh(x)$, integrals [5,19,20]. The fuzzy integrals of h over $\boldsymbol{X}$ with respect to g_λ are defined as:

$$\begin{aligned} Su(x) &= \bigvee_i \left[h(x_i) \wedge g_\lambda(\boldsymbol{X}_i) \right] \\ Ch(x) &= \bigvee_i \left[\left(h(x_i) - h(x_{i-1}) \right) g_\lambda(\boldsymbol{X}_i) \right] \\ Sh(x) &= \bigvee_i \left[h(x_i) \cdot g_\lambda(\boldsymbol{X}_i) \right] \end{aligned} \tag{12}$$

where $\boldsymbol{X}_i=\{x_1, x_2, \ldots, x_i\}$ and $h(x_0)=0$. While several possible interpretations exist for the conceptual meaning of a fuzzy integral [46,48], in this discussion it is considered to mean the maximum degree of belief (for a prediction or classification outcome) obtained by the fusion (aggregation) of several sources of objective evidence.

Integrating the results from multiple classifiers involves using their respective confusion matrices to compute the fuzzy densities for each of the classifiers in order to determine the fuzzy measures used in (12). To this end, the technique described in [4] is followed primarily and is briefly described here. Let $\boldsymbol{R}_k=(n_{kij})$ be the $c \times c$ confusion matrix for classifier, k, where n_{kii} is the number of class i patterns that were correctly classified by k and n_{kij} $(i \neq j)$ is the number of class i patterns that were incorrectly assigned to class j by k. The preliminary fuzzy density of class i with respect to classifier k, $0 < g^*_{ki} < 1$, is

$$g^*_{ki} = \frac{n_{kii}}{\sum_{j=1}^{c} n_{kij}} \tag{13}$$

These densities must be adjusted to take into account the frequencies of correct and incorrect classifications within and across the set of classifiers. This leads to the following expressions

$$\delta_{kij} = \begin{cases} 1 & i = j \\ \dfrac{n_{kii} - n_{kij}}{n_{kii}} & i \neq j \\ \varepsilon & n_{kii} < n_{kij} \end{cases}, \gamma_{kij} = \begin{cases} 1 & n_{kij} < n_{lij} \\ \dfrac{n_{lij}}{n_{kij}} & n_{kij} \geq n_{lij} \\ \varepsilon & n_{kij} = 0 \end{cases} \tag{14}$$

where ε is a small positive constant. The corrected fuzzy density, g_{ki}, may now be computed as

$$g_{ki} = g_{ki}^{*} \times \left(\delta_{kir} \times \cdots \times \delta_{kis}\right)^{w_1} \times \left(\gamma_{kir} \times \cdots \times \gamma_{kis}\right)^{w_2} \tag{15}$$

where w_1 and w_2, (w_1+w_2=1) are weighting factors and r and s are the indices of those pattern classes for which classifier k produced the highest classification accuracy. The first adjustment, $\delta \in (0,1]$, reflects the pattern misclassifications within the confusion matrix for k. As the pattern misclassifications increase, $\delta \to 0$ (the third condition in (14) represents the degenerate case when more patterns of a particular class are misclassified than correctly classified). The second adjustment, $\gamma \in (0,1]$, reflects the pattern misclassifications across all classifiers with respect to k. As the pattern misclassifications increase, $\gamma \to 0$ (the third condition in (14) is the degenerate case when no patterns of a particular class are correctly classified).

Finally, the Sugeno, Choquet, and Shilkret integrals can exploit several variants of h including: $h_c(x)$, contrast intensification as defined by (11), and $h_p(x)=x^p$ (p>0), where $x \in [0,1]$ is the classifier's predicted class label assignment. When 0<x<1, $h(x)$ will act to dilate membership values, while concentration will occur when x>1. In order to constrain the number of parameters, the standard fuzzy set based definitions for concentration (p=2) and dilation (p=0.5) are normally used. In total, four variants are typical candidates for the integrals: $h_c(x)$, $h_{0.5}(x)$, $h_2(x)$, and $h_1(x)$ (identity). Finally, using equations (12)–(15), the actual class label output from the set of pattern classifiers is the one with the highest integrated value.

7 Experiments, Analysis and Results

In this concluding section, we present a series of experiments, which employed the classification approaches described above, relating to the interpretation, analysis, and classification of several biomedical datasets. A summary of the results, listed in Table 2, may be found at the end of this section.

7.1 FQE

In [34], MR spectra were obtained (360 MHz) for 25 thyroid biopsies: 16 papillary carcinomas and 9 normal. Two spectral regions were analyzed: the main lipid CH2 and CH3 peaks, 0.64–2.59 ppm; and the choline-like species, 2.59–3.41 ppm. Analysis was based on 170 features for the choline region and 400 features for the lipid region. As a benchmark, the inputs to an MLP classifier were the 10 principal components of the dataset that accounted for 97% of the cumulative variance [37]. FQE was used with 680 (choline) and 1600 (lipid) membership values.

FQE significantly outperformed the benchmark: P_O=0.92 versus P_O=0.64 (choline); P_O=0.88 versus P_O=0.80 (lipid). Of particular interest is the significant reduction in convergence rate for the FQE MLP, $O(10^3)$ versus $O(10^6)$ for the benchmark.

In [30], data were analyzed pertaining to tonsillectomy/adenoidectomy patients with predispositions to excessive bleeding. These blood abnormalities include hemophilia, a hereditary hemorrhagic diathesis due to coagulation cofactor FVIII deficiency; von Willebrand's disease, a diathesis associated with von Willebrand protein antigen factor deficiencies or in the activity measured as the restocetin cofactor; and thrombopathy, a platelet function defect measured as the occurrence of at least two abnormal platelet aggregation [21]. Data were collected from the patient database associated with a hematology expert system containing information relating to coagulation laboratory test results and patients responses to a bleeding tendency questionnaire.

Two major experiments were conducted. In the first, 96 patient records (patterns) were assigned to one of three disease states (class labels): 42 hemophilia (H), 30 platelet function defect (P), and 24 von Willebrand's disease (V). LDA, MLP, and FQE (with MLP) classifiers were used in the analysis: respectively, κ=0.55 (moderate agreement), κ=0.71 (substantial agreement), and κ=0.79 (substantial agreement). MLP and FQE also had consistently better classification results across all three disease states with particularly strong improvements with H and V. While FQE outperformed MLP with respect to correctly classifying P (80% versus 70%), it under performed with respect to V (83% versus 88%). However, FQE was clearly superior in classifying H; 93% versus 81%. FQE, on average, converged 4.2 times faster during the training phase than MLP.

In the second set of experiments, a different gold standard was used (derived from the expert system) to assign 191 patient records to either a normal (N) or abnormal (A) class. The records were randomly assigned to a design set (60 N and 60 A) or a validation set (42 N and 29 A). The respective κ scores for LDA, MLP, and FQE were 0.16 (slight agreement), 0.39 (fair agreement), and 0.46 (moderate agreement).

In [32], dispersion-adjusted FQE (DFQ) MLP classifiers were used in the analysis and classification of three biomedical datasets found in the Machine Learning Repository (http://mlearn.ics.uci.edu/MLSummary.html) at the University of California, Irvine. The patterns in these three datasets belong to one of two possible classes: "target", where the pattern belongs to an abnormal or disease state; and, "control", where the pattern belongs to a normal or control state.

In the first case, the heart data [16] is a description of diagnoses relating to N=267 cardiac single proton emission computed tomography images [10]. The n=44 features relate to frequency information across 22 different regions of interest and alternate between images taken while the patient was at rest or during a controlled stress condition (target=55, control=22). The overall classification accuracy using the original features was P_O=0.80 while the FQE accuracy was P_O=0.92 and the DFQ accuracy was P_O=0.95 (a respective 15% and 19% increase in performance). DFQ decreased the false positive error rate from 10% to 7% with an overall increase in accuracy of 3%.

In the second case, each of the N=155 patterns (target=32, control=123) within the hepatitis dataset [2] is composed of 19 features: 6 nominal features (age, bilirubin, alkaline phosphate, SGOT, albumin, and protime) and 13 binary features (sex, steroids, antivirals, fatigue, malaise, anorexia, large liver, firm liver, palpable spleen, spiders,

ascites, varices, and histology). The overall accuracy using the original features was P_O=0.88 while the FQE accuracy was P_O=0.91 and the DFQ accuracy was P_O=0.94 (a respective 3% and 7% increase in classifier performance). With FQE, this improvement was gained exclusively by a reduction in the false negative error rate (from 37% to 22%). In the DFQ case, a greater reduction in the false negative error rate (19%) was achieved with an overall increase in accuracy of 3% compared to the FQE encoding.

In the third case, the lung cancer data [12], which comprises 56 nominal features taking on integer values (0–3), represents three different types of pathological lung cancers. Due to the paucity of patterns (N=32) in this dataset, and in the interest of simplifying the comparative analysis with the other two biomedical datasets, the two classes with the fewest patterns are merged into one pathological (target) case (control=13, target=19). The overall classification accuracy using the original features was P_O=0.63 while the FQE accuracy was P_O=0.78 and the DFQ accuracy was P_O=0.84 (a respective 23% and 33% increase in classifier performance). DFQ achieved an 8% improvement in classification performance compared to FQE.

7.2 GSA

In [31], GSA was used in the analysis of a biomedical dataset composed of 206 ^{1}H MR spectra (360 MHz, 37°) consisting of 95 meningiomas (M), 74 astrocytomas (A), and 37 control samples of non-tumorous brain tissue from patients with epilepsy (E). The biomedical spectra (n=550 in the region of 0.3–4.0 ppm) were randomly assigned to either a design (N_D=80, with 29 M, 31 A, and 20 E) or a validation set (N_V=126). Applying GSA to the gold standard (provided by a pathologist) improved the overall diagnostic (classification) performance of an MLP classifier by 13%: κ=0.80 versus κ=0.71 using the original design class labels.

Although none of the spectra (patterns) in the validation set was reclassified, two validation spectra were flagged as outliers (two M spectra were flagged as A), and these spectra were indeed misclassified as A. Classification errors were also more conservative. Using the original class labels, 5 E's (control) were classified as tumors (M or A) and 4 tumors as control. However, in the case of GSA, only 1 E was misclassified as a tumor while only 3 tumors were misclassified as control.

7.3 SFS

In [33], SFS was used in the analysis and classification of N=444 ^{1}H MR spectra (360 MHz at 37°C) of isolates of five different species of Candida yeast (n=1500): 104 albicans (A), 93 parapsilosis (P), 81 krusei (K), 75 tropicalis (T), and 91 glabrata (G). The design set comprised 50 randomly selected patterns from each class. The feature region cardinality range was 7–1231.

The mean accuracy (for the 10,000 MLP processes) was P_O=0.83 for the validation set. The best accuracy score was P_O=0.95 for an MLP using only 16 of the 1500 features (~1%). Interestingly, the top four accuracy scores were achieved by MLPs that used less than 20 features.

In [29], SFS was used in the classification of two biomedical datasets. In the first case, 186 infrared spectra of synovial joint fluid were assigned to one of three disease

states: 72 rheumatoid arthritis (R), 72 osteo-arthritis (O), and 42 control samples (C). The spectra (n=2801) cover the wavelength range 1000–3700 cm^{-1}. The pattern design set contained 28 randomly selected spectra from each class. The feature region cardinality range was 7–192 and the mean κ score was 0.79 for the 20,000 MLP and PNN classification processes. The best validation κ score was 0.86±0.02 (almost perfect agreement) using the MLP classifier with only 20 of the 2801 original features (<1%). The best validation κ score for PNN was 0.84±0.02 (almost perfect agreement) using 25 of the original features (<1%). Using the original 2801 spectral features (i.e., no stochastic feature selection), the PNN benchmark produced a validation set κ score of only 0.51 (moderate agreement), while the MLP benchmark κ score was only 0.29 (fair agreement).

In the second case, this is likely due to over-fitting of the design (training) data as is evidenced by the high κ score of 0.88. Due to the inversion of the large covariance matrix, LDA produced spurious results. Next, the original infrared spectra were averaged down to 100 features. All three benchmarks performed well (substantial agreement): PNN, κ=0.69; MLP, κ=0.74; LDA, κ=0.69. While slightly worse than the average of all 10000 PNN (and MLP) runs using SFS, they were appreciably worse than the best runs. In the second case, 227 MR spectra of a biological fluid discretized (n=512) were assigned to one of three classes: 108 normal (N), 54 of borderline character (B), and 65 abnormal (A). The design set contained 36 randomly selected samples from each class.

The feature region cardinality range was 4–212. For the 10000 MLP classification processes, the mean κ score was 0.42 for the validation set. For the 10000 PNN classification processes, the mean κ score was 0.48 for the validation set. The best validation κ score was 0.54±0.02 (moderate agreement) using MLP and 83 of the 512 original features (16%). The best validation κ score for PNN was 0.48±0.02 (moderate agreement) using 79 of the original features (15%). Using the original 512 spectral features (again no feature selection), all benchmarks performed poorly (only fair agreement): PNN, κ=0.38; MLP, κ=0.25; LDA, κ=0.24. As with the infrared spectra, the MLP likely over-fitted the design data (κ=0.88). Finally, the original dataset was averaged down to 128 features. All benchmarks had moderate levels of agreement: PNN, κ=0.47; MLP, κ=0.46; LDA, κ=0.43.

7.4 Classifier Aggregation

In [35], fuzzy aggregation was used in conjunction with SFS in the analysis and classification of N=191 MR spectra (n=3380) of a biofluid that were assigned to one of two classes by a medical expert: 116 normal and 75 abnormal. The design set comprised 58 randomly selected patterns from each class. Three transformed dataset variants were generated: first derivative; rank ordered; and first derivative with rank ordering. The best validation set result was P_O=0.79 using the fuzzy aggregation approach with rank ordered transformed features, which is an 8% improvement over the corresponding best individual (PNN) classifier. Further, the aggregated approach outperformed the corresponding best individual classifiers across all variants: respectively (P_O), 0.76/0.74, 0.74/0.62, 0.79/0.73, 0.75/0.73.

Table 2. Summary of biomedical data classification results

Description	N	n	P	Method	Benchmark
FQE: MRS Thyroid I	25	170	P_O	0.92	0.64
FQE: MRS Thyroid II	25	400	P_O	0.88	0.80
FQE: Hematology I	96	11	κ	0.79	0.71
FQE: Hematology II	191	9	κ	0.46	0.39
DFQ: Heart	267	44	P_O	0.95	0.92
DFQ: Hepatitis	155	19	P_O	0.94	0.91
DFQ: Lung Cancer	32	56	P_O	0.84	0.78
GSA: MRS Brain	206	550	κ	0.80	0.71
SFS: MRS Candida	444	1500	P_O	0.95	0.83
SFS: Synovial Fluid	186	2801	κ	0.86	0.74
SFS: MRS Biofluid	227	512	κ	0.54	0.47
Fusion: MRS Biofluid	191	3380	P_O	0.79	0.73

Each entry lists the biomedical classification method examined (as described in this section), the dataset used in the evaluation, the number of patterns (N), the number of features (n), the performance measure (P), the method's overall accuracy, and the accuracy for the best benchmark.

8 Conclusion

The analysis, interpretation, and classification of biomedical data are replete with pattern recognition challenges stemming from the curse of dimensionality and tarnished gold standards. This chapter presents a computational intelligence based methodology, which remediates these challenges, exploiting strategies and methods inspired by the granular computing paradigm. Stochastic feature selection, gold standard class label adjustment, classifier aggregation, and fuzzy quantile encoding may be used singly or in concert within a classification system. As pre- and post-processing approaches, they may easily be incorporated into investigators' classifiers of choice.

Acknowledgments. We thank Conrad Wiebe and Aleksander Demko for implementing the stochastic feature selection algorithm and associated libraries. The following researchers are gratefully acknowledged for making their respective datasets publicly available for use by their peers: K.J. Cios and L.A. Kurgan for the SPECTF heart data; G. Gong and B. Cestnik for the hepatitis data; S. Aeberhard for the lung cancer data; and, U. Himmelreich for the yeast data. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Bishop, C.M.: Neural networks and their applications. Rev. Sci. Instrum. 65, 1803–1832 (1994)
2. Cestnik, B., Kononenko, I., Bratko, I.: ASSISTANT 86: A knowledge elicitation tool for sophisticated users. In: Bratko, I., Lavrac, N. (eds.) Progress in Machine Learning. Sigma Press, Wilmslow (1987)

3. Cheng, B., Titterington, D.M.: Neural networks: a review from a statistical perspective. Stat. Sci. 9, 2–54 (1994)
4. Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition. World Scientific, New Jersey (1996)
5. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier 5, 131–295 (1953)
6. Demko, A.B., Pizzi, N.J., Somorjai, R.L.: Scopira – A system for the analysis of biomedical data. In: Proc. IEEE Can. Conf. Electr. Comput. Eng., Winnipeg, Canada, May 12–15, 2002, pp. 1093–1098 (2002)
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley Interscience, New York (2000)
8. Everitt, B.S.: Moments of the statistics kappa and weighted kappa. Br. J. Math. Stat. Psychol. 21, 97–103 (1968)
9. Fleiss, J.L.: Measuring agreement between judges on the presence or absence of a trait. Biom. 31, 651–659 (1975)
10. Friebolin, H.: Basic One- and Two-Dimensional NMR Spectroscopy. Wiley & Sons, New York (1998)
11. Grabish, M., Murofushi, T., Sugeno, M.: Fuzzy measure of fuzzy events defined by fuzzy integrals. Fuzzy Sets Syst. 50, 293–313 (1992)
12. Hong, Z.Q., Yang, J.Y.: Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognit. 24, 317–324 (1991)
13. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101 (1964)
14. Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainty, and Information. Prentice-Hall, Englewood Cliffs (1988)
15. Kuncheva, L.I., Steimann, F.: Fuzzy diagnosis. Artif. Intell. Med. 16, 121–128 (1999)
16. Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M., Goodenday, L.S.: Knowledge discovery approach to automated cardiac SPECT diagnosis. Artif. Intell. Med. 23, 149–169 (2001)
17. Landis, J.R., Koch, G.G.: The measurements of observer agreement for categorical data. Biom. 33, 159–174 (1997)
18. Liu, H., Hussain, F., Tan, C.L., Dash, M.: Discretization: An enabling technique. Data Min. Knowl. Discovery 6, 393–423 (2002)
19. Mesiar, R., Mesiarová, A.: Fuzzy Integrals—What Are They? Int. J. Intell. Syst. 23, 199–212 (2008)
20. Murofushi, T., Sugeno, M.: An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure. Fuzzy Sets Syst. 29, 201–227 (1989)
21. Nosek-Cenkowska, B., Cheang, M.S., Pizzi, N.J., Israels, E.D., Gerrard, J.M.: Bleeding/bruising symptomatology in children with and without bleeding disorders. Thromb. Haemost. 65, 237–241 (1991)
22. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. 33, 1065–1076 (1962)
23. Pavia, D.L., Lampman, G.M., Kriz, G.S.: Introduction to Spectroscopy. Harcourt Brace College, Fort Worth (1996)
24. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. 177, 3–27 (2007)
25. Pedrycz, W.: Granular Computing: The Emerging Paradigm. J. Uncertain Syst. 1, 38–61 (2007)
26. Pedrycz, W.: Granular Computing: An Emerging Paradigm. Springer, Heidelberg (2001)
27. Pedrycz, W., Gomide, F.: Fuzzy Systems Engineering: Toward Human-Centric Computing. Wiley & Sons, New York (2007)
28. Pizzi, N.J.: Biomedical data analysis using dispersion-adjusted fuzzy quantile encoding. In: Proc. Annu Meet North Am. Fuzzy Inf. Process Soc., New York, USA, #50010 (6 pages), May 19–22 (2008)
29. Pizzi, N.J.: Classification of biomedical spectra using stochastic feature selection. Neural Netw. World 15, 257–268 (2005)

30. Pizzi, N.J.: Bleeding predisposition assessments in tonsillectomy/adenoidectomy patients using fuzzy interquartile encoded neural networks. Artif. Intell. Med. 21, 65–90 (2001)
31. Pizzi, N.J.: Fuzzy preprocessing of gold standards as applied to biomedical spectra classification. Artif. Intell. Med. 16, 171–182 (1999)
32. Pizzi, N.J., Pedrycz, W.: An analysis of potentially imprecise class labels using a fuzzy similarity measure. In: Proc. World Congr. Comput. Intell., Hong Kong, June 1–6, pp. 667–672 (2008)
33. Pizzi, N.J., Pedrycz, W.: Classification of magnetic resonance spectra using parallel randomized feature selection. In: Proc. Int. Jt. Conf. Neural Netw., Budapest, Hungary, July 25–29, 2004, pp. 2455–2460 (2004)
34. Pizzi, N., Somorjai, R.L.: Fuzzy encoding as a preprocessing method for artificial neural networks. In: Proc. World Congr. Neural Netw., San Diego, USA, June 5-9, 1994, pp. 643–648 (1994)
35. Pizzi, N.J., Wiebe, C., Pedrycz, W.: Biomedical spectral classification using stochastic feature selection and fuzzy aggregation. In: Proc. Ann. Meet North Am. Fuzzy Inf. Process Soc., San Diego, USA, June 24–27, 2007, pp. 360–365 (2007)
36. Pizzi, N.J., Alexiuk, M.D., Pedrycz, W.: Classification of biomedical spectra using fuzzy interquartile encoding and stochastic feature selection. In: Proc. IEEE Symp. Ser. Comput. Intell. Data Min., Honolulu, USA, June 1–6, pp. 668–673 (2007)
37. Pizzi, N., Choo, L.-P., Mansfield, J., Jackson, M., Halliday, W.C., Mantsch, H.H., Somorjai, R.L.: Neural network classification of infrared spectra of control and Alzheimer's diseased tissue. Artif. Intell. Med. 7, 67–79 (1995)
38. Ripley, B.D.: Neural networks and related methods for classification. J. Royal Stat. Soc. [B] 56, 409–456 (1994)
39. Rumelhart, D.E., McClelland, J.L.: Parallel Distributed Processing, vol. 1. MIT Press, Cambridge (1986)
40. Seber, G.: Multivariate Observations. Wiley & Sons, New York (1984)
41. Snir, M., Gropp, W.: MPI: The Complete Reference. MIT Press, Cambridge (1998)
42. Somorjai, R.L., Dolenko, B., Nikulin, A.K., Pizzi, N., Scarth, G., Zhilkin, P., Halliday, W., Fewer, D., Hill, N., Ross, I., West, M., Smith, I.C.P., Donnelly, S.M., Kuesel, A.C., Briere, K.M.: Classification of 1H MR spectra of human brain neoplasms: The influence of preprocessing and computerized consensus diagnosis on classification accuracy. J. Magn. Reson. Imaging 6, 437–444 (1996)
43. Specht, D.F.: Probabilistic neural networks. Neural Netw. 3, 109–118 (1990)
44. Sugeno, A.: Fuzzy measures and fuzzy integrals: A survey. In: Gupta, M.M., Saridis, G.N., Gaines, B.R. (eds.) Fuzzy Automata and Decision Processes, pp. 90–102. North Holland, Amsterdam (1977)
45. Sugeno, A.: Theory of fuzzy integral and its applications. PhD Thesis, Tokyo Institute of Technology (1972)
46. Tahani, H., Keller, J.M.: Information fusion in computer vision using the fuzzy integral. IEEE Trans. Syst. Man Cybern. 20, 733–741 (1990)
47. Vapnik, V.: Statistical Learning Theory. Wiley & Sons, New York (1998)
48. Weiss, S.M., Kulikowski, C.A.: Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems. Morgan Kaufmann, San Mateo (1991)
49. Witten, I.H., Eibe, F.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Mateo (2005)
50. Zadeh, L.A.: Fuzzy logic = Computing with words. IEEE Trans. Fuzzy Syst. 4, 103–111 (1996)
51. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. 3, 28–44 (1973)