Fuzzy Set Similarity for Feature Selection in Classification

Valerie Cross
Computer Science and Software Engineering
Miami University
Oxford, OH USA
crossv@miamioh.edu

Michael Zmuda
Computer Science and Software Engineering
Miami University
Oxford, OH USA
zmudam@miamioh.edu

Rahul Paul
Computer Science and Engineering
University of South Florida
Tampa, FL USA
rahulp@mail.usf.edu

Lawrence Hall
Computer Science and Engineering
University of South Florida
Tampa, FL USA
lohall@mail.usf.edu

***Abstract*—A problem for machine learning research occurs when many possible features exist but the training data examples are very few. For example, microarray data typically have a much larger number of features, the genes, as compared to the number of training data examples, the patients. One approach is to first determine the best features for prediction and then to group features based on a measure of their relatedness. The concordance correlation coefficient has been used to place somewhat correlated features into disjoint groups of similar features. Multiple base classifiers are created by randomly picking one feature from each of the feature groups and then the collection of base classifiers is used in an ensemble classifier. Each classifier in the ensemble provides a vote. The majority vote is used to produce the final class prediction. This paper investigates grouping features using fuzzy set similarity measures as well as the concordance correlation coefficient as a relatedness measure. The performance of these different measures is compared in terms of accuracy, sensitivity, specificity, and F-measure using the ensemble classifiers created with the different relatedness measures. Four microarray gene expression data sets are used in the experiments to determine the usefulness of fuzzy set similarity measures and how they compare with the concordance correlation coefficient. Using the concordance correlation coefficient to guide clustering is not superior to fuzzy set similarity measures. Depending on the particular data set and performance measure being used, different fuzzy set similarity measures perform better than or just as well as the concordance correlation coefficient.**

Keywords—feature selection, fuzzy set similarity measures, concordance correlation coefficient, microarray data.

I. Introduction

The success of machine learning algorithms is typically reliant on the quality of the machine readable data that they work on. Quality factors include whether there is irrelevant, redundant, unreliable or noisy data. Another factor for machine learning algorithms is the high-dimensional datasets becoming more and more significant in research. A challenge is to find a set of features, reduced as much as possible, that are able to accurately classify the sample data. Because of this challenge, feature selection has become more indispensable than ever to achieve dimensionality reduction [1].

Dimensionality reduction is very essential in biological applications such as DNA-microarrays and proteomics since these applications commonly have high dimensionality with a small number of examples. Biomedical researchers need a small set of highly discriminatory features for which they will then invest substantial time and research effort. Since feature selection also preserves the original features, it is particularly useful for applications that require the original features for model interpretation and knowledge extraction. Consequently, selecting the best set of differentiating features for classification in biological applications has received attention in data mining and machine learning research [2][3].

The numerous algorithms for feature selection have been categorized into filter, wrapper and sparsity-based approaches [4]. The research conducted for this paper uses a filter method to rank the features using their relevance as determined by statistical measures taken over the underlying data characteristics. These top-ranked features are then used to build a single base classifier. Multiple single base classifiers are created and then aggregated into an ensemble. Research suggests that an ensemble created from multiple base classifiers can improve both performance and confidence in the results. The ensemble in this research aggregates the predictions from the base classifiers using simple majority voting [5].

The base classifiers for an ensemble can be created in different ways, but the approach taken here is similar to the random subspace method (RSM) [6], which trains the same base classifier by random sampling of features from the feature space. RSM, however, assumes that the features randomly selected are not highly correlated. This assumption can affect its performance due to feature redundancy and efficiency. A modified version of RSM [7] takes into consideration that some features might be related as measured by the concordance correlation coefficient [8]. This method is referred to as the concordance correlation coefficient based random subspace method (CCC_RSM). In this approach, the features are grouped into feature subsets based on their CCC values and the random sampling then occurs within a feature subset. For example, if 10 feature subsets result from grouping using the CCC, then 10 features, one randomly selected from each feature subset, are used to train a base classifier. The research presented in this paper investigates the use of other measures of relatedness between the features of the classification problem, specifically, fuzzy set similarity measures, to group the features into feature subsets. The details are described in Section III below.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

The paper organization is as follows: Section II reviews the relatedness measures used in this research including the concordance correlation coefficient and the fuzzy set similarity measures. Section III describes the experimental design and its parameters as well as the data sets used to evaluate the different measures. Section IV compares the results from applying these measures with respect to the feature subsets formed and their performance in the classification task. Finally, Section V presents the conclusions and possible future work.

II. Relatedness Measures

In [7] the concordance correlation coefficient (CCC) was used as the relatedness measure between the features in order to group features together for the selection step. The random subspace method [6] does not use relatedness grouping. With grouped features, a base classifier is repeatedly trained using a set of features where each feature is randomly selected from each of the groups. If G groups are formed using a relatedness measure, then G features are in the feature set used to train the base classifier.

To summarize, instead of randomly selecting the features from the original feature space, a relatedness measure is used to form the groups of related features. This research investigates the effects of using different fuzzy set similarity measures to determine relatedness in creating the groups. First, a description of how the patient microarray data is represented as fuzzy sets is provided. Then the CCC and several different fuzzy set similarity measures are described.

A. Creation of Fuzzy Sets Representing Features

For purposes of the applying fuzzy set similarity measures to determine agreement between two different features, each gene feature is represented as a fuzzy set over the patients as provided in the sample data sets. The gene expression levels must be normalized to specify a degree of membership in [0, 1]. The patient's membership degree in the fuzzy set specifies the level of gene expression for that patient's microarray data.

B. Concordance Correlation Coefficient

The concordance correlation coefficient (CCC) measure is a bivariate relationship in terms of agreement between two values [8]. This measure differs from the Pearson correlation which measures how much the relationship is linear. If a scatterplot of the pair of points is examined, high concordance correlation occurs when the scatterplot points are close to the 45 degrees line of perfect concordance which runs diagonally on the scatterplot. High Pearson correlation occurs when the scatter points are near any straight line. CCC values lie in the interval [-1, 1] where -1 implies negative agreement and +1 positive agreement. A zero value indicates no agreement.

The CCC for two variables A and B uses the means, variances (var) and covariance (cov) of A and B. It is defined as

$$\mathrm{CCC}(A, B) = \frac{2 * \mathrm{cov}(A, B)}{\mathrm{var}(A) + \mathrm{var}(B) + (\overline{A} - \overline{B})^2} \quad (1)$$

Other correlation coefficients may be used, but since the CCC has recently been used as a relatedness measure for grouping features for classification, it is used as one of the relatedness measures in this current research.

C. Zaheh's Sup-Min Partial Matching

In [9] a detailed and thorough review of a variety of fuzzy set similarity measures is provided. Zadeh's consistency index also known as the sup-min or partial matching index falls into the set-theoretic category of fuzzy similarity measures. It roughly estimates the similarity between two fuzzy sets by finding at what domain values they intersect and determines their similarity by taking the highest membership degree among their intersection points. Given two fuzzy sets A and B, similarity between the two is determined as

$$S_{\mathrm{Zadeh}}(A, B) = \sup_{u \in U} T(A(u), B(u)) \quad (1)$$

where T can be any t-norm. Usually the minimum is used for the t-norm, as done in this work. This measure only provides a rough calculation for the similarity value between the two fuzzy sets.

D. Jaccard

The fuzzy Jaccard similarity measure is defined as a fuzzy extension of the Jaccard index [10] between two crisp sets by replacing set cardinality with fuzzy set cardinality. This fuzzy set similarity measure is also in the set theoretic category but provides a more comprehensive view of similarity between the two fuzzy sets since all elements in both fuzzy sets are taken into account not just the intersection point as in sup-min. Given two fuzzy sets A and B, similarity between the two is defined as

$$S_{\mathrm{Jaccard}}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

so the similarity is measured by the proportion of the area of the intersection of the two fuzzy sets to the area of the union of the two fuzzy sets. In this work the minimum operator is used for intersection and the maximum operator is used for union of two fuzzy sets.

E. Aggregation of Inclusion Measures

Another way to create a fuzzy set similarity measure is to aggregate a fuzzy inclusion measure using a symmetric aggregation operator [9]. A fuzzy inclusion measure determines how much of one fuzzy set is included in another fuzzy set. For example, the inclusion of fuzzy set A in fuzzy set B is specified as

$$S_{\mathrm{Inc}}(A, B) = \frac{|A \cap B|}{|A|} \quad (3)$$

A fuzzy similarity measure is then created as

$$S_{\mathrm{IncAgg}}(A, B) = \mathrm{agg}[S_{\mathrm{Inc}}(A, B), S_{\mathrm{Inc}}(B, A)] \quad (4)$$

where agg is the average, minimum or maximum aggregation operator.

F. Cosine Measure

The cosine similarity measure between two fuzzy sets A and B is represented as

$$S_{cos}(A, B) = \frac{A \circ B}{\|A\| \, \|B\|}. \quad (5)$$

This measure views each feature vector as a vector in *n* dimensional space and computes the cosine of the angle between the two vectors. When the two vectors are coincident, the cosine is 1; therefore, the similarity is 1. When the vectors are perpendicular, the cosine is 0; therefore, there is no similarity. The data used in this work contain non-negative values and so it is not possible for the two vectors to form an obtuse angle, therefore, $S_{cos} \geq 0$.

TABLE I MICROARRAY DATASETS

Dataset name	No. of features	No. of samples	Pos/ Neg	Download
Breast	7129	44	21/22	https://www.csie.ntu.edu.tw/_cjlin/libsvmtools/datasets/binary.html#breast-cancer
CNS	7129	60	39/21	http://csse.szu.edu.cn/staff/zhuzx/Datasets.Html
Colon	2000	62	40/22	http://microarray.princeton.edu/oncology/affydata/index.html
Leukemia	7129	72	47/25	http://csse.szu.edu.cn/staff/zhuzx/Datasets.Html

III. EXPERIMENTAL DESIGN AND DATA SETS

The research objective is to compare how the different relatedness measures perform in grouping features to create an ensemble classifier. The results are compared and analyzed to determine the effects of the measures on the classification results. To perform this analysis, a systematic series of machine learning experiments were conducted, with the main control variable being the similarity measure used to form the feature groups. The data sets used in these experiments are provided in Table I.

Ultimately, we adopted using leave-one-out cross validation [11]. Central to this procedure is the repeated application of the learning process on a training set and test item. As seen in Fig. 1, the first step is to normalize the data to ensure that the values are in the interval [0, 1] so that it is suitable for fuzzy set similarity measures. The leftmost data column shows the original 4 hypothetical training instances and 1 test instance. As one goes left to right in Fig. 1 on the data, each of the training set's 3 feature vectors is normalized by finding the min and max in each column independently. In the example, the three feature min/max pairs are: 0/4, 0/100, and 10/110. Normalizing these samples linearly using (v – min) / (max – min) results in the second column. Since the test sample cannot be used at any point in the learning process (including normalization), we store the min/max pairs of the training set and use them to scale the test sample, but also clamp any values that fall outside the range [0, 1]. In the example, the third feature of the test sample is clamped back into the range [0, 1].

After normalization, the ReliefF procedure which is shown below the normalization process in Fig. 1 is used to rank order the most useful features [12]. In the example, the rankings, from best to worst, are F1, F3, and F2. A value N denotes how many features are to be kept. In this case, N=2 and therefore feature 2 is discarded since it has the lowest ranking. The last column in the Relief Ranking procedure in Fig. 1 shows the resulting data that is used during the one cycle of leave-one-out cross validation.

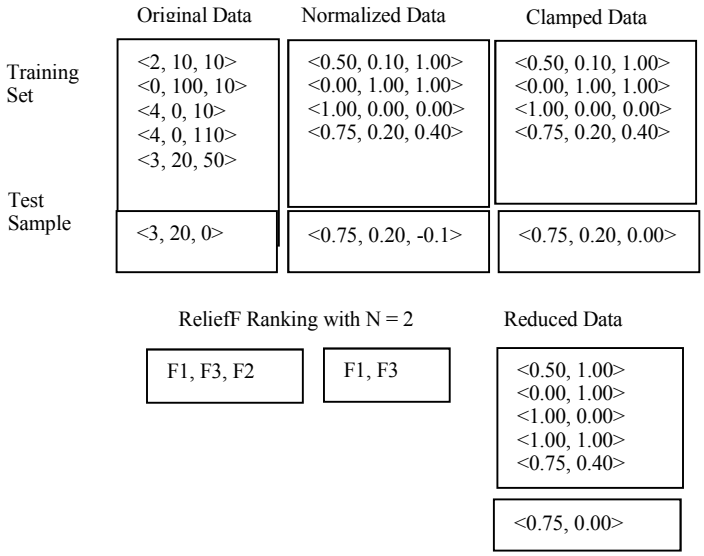

Fig. 1. Normalization and Feature Reduction

The ReliefF procedure evaluates each feature in isolation and does not consider the combined effects of what features work well together. Thus, if a particularly discriminative feature were to be replicated several times, ReliefF would rate all of those copies highly, even though it is beneficial to assemble a cooperative set of feature vectors. Nonetheless, ReliefF does provide a good way to assess the merit of individual features, despite its inability to assess the merit of a *collection* of features. This latter deficiency can be mitigated by increasing the number of top features selected, N. Of course, arbitrarily increasing N renders ReliefF useless and using all features becomes the standard. The nature of the data may, however, suggest there may be values of N that may reduce the number of features, yet provide enough diversity to ensure that the resulting set of features are able to perform effective classification.

The individual feature vectors (i.e., the genes' normalized values for all patients) of the reduced data are then placed into G disjoint groups, where each group is designed to contain similar features. This process is done using hierarchical clustering [13]. Classical clustering uses the concept of *distance* between individual items. Distance is essentially the opposite of similarity. So, distance is defined as 1.0 – similarity, when doing clustering.

The clustering procedure starts with each feature vector residing in its own cluster. Then, the two closest clusters are merged into one new, larger, cluster and the process repeated. Several approaches exist for determining the distance between clusters containing multiple items: 1) distance between the two clusters' centroids 2) distance between the closest pair of elements in the two clusters or 3) distance between the farthest pair of elements in the two clusters. In this work, option 3 is adopted. This approach tends to merge clusters that are relatively compact and are closely situated. Several options exist to determine when the merging should stop. In this work, clusters are continually merged until a pre-specified number of clusters is obtained.

The clustering procedure starts with the normalized, reduced, training data. First, the similarity value of each feature-pair is computed by the similarity measure being used. This value is then converted to a distance and stored in a symmetric distance matrix. This distance matrix is then consulted during the hierarchical clustering procedure previously described to obtain G groups of disjoint feature sets.

Algorithm 1.

```
1.  LOO-CV(Set<Instance> samples, int N, int G, Similarity S, int E):
2.    For each (Instance testSample in samples):
3.      training = samples – { testSample }
4.      Normalize feature vectors in training and testSample
5.      Use ReliefF to select N features from training (and
          testSample)
6.      Cluster feature vectors IDs of training into G groups using S
7.      ensemble = { }
8.      Repeat E times
9.        Randomly select one feature from each cluster,
            reducing the training set and testSample
            to G features. Store the reduced testSample for
            later testing
10.       ensemble = ensemble ∪ Learn J48 classifier using
            reduced training set
11.     Apply each classifier in ensemble to testSample, with only
          the proper features selected.
          Record if ensemble correctly classifies testSample or
          the type of classification error.
```

Fig. 2. Normalization and Feature Reduction

Creating the distance matrix requires the relatedness value of all pairs of the G feature vectors, where the number of pairs is O(G^2). Computing the relatedness of one pair of feature vectors is O(M), where M is the number of samples in the training set. Preparing the distance matrix is, therefore, O(MG^2) with M and G typically being small.

Algorithm 1 in Fig. 2 describes the leave-one-out validation process. It uses the procedures described to form an ensemble consisting of E decision tree classifiers. Weka's J48 decision tree classifier [14] is used with default parameters. Line 2 cycles through all instances, one at a time. Lines 4-6 normalize, reduce, and perform clustering described in Fig. 1. Lines 7-10 form the ensemble of J48 classifiers. To form the training feature set ultimately used in J48 learning, a feature is randomly selected from each of the G groups. It is worth noting that this is the second instance where features have been down selected in the entire algorithm.

The new feature selector is similar to the one constructed in [12] since this feature selector combines ranking-based methods using ReliefF (Line 5), grouping-based methods with its use of relatedness measures (Line 6), and random subspace methods (Line 9) since it randomly selects a feature from each group of related features. It differs in that it can vary the relatedness measure S used to group the features and its grouping method uses a standard clustering approach previously described. The number of groups or clusters are varied instead of varying the threshold value of the CCC required to be placed in the same group as done in [7]. Both approaches determine a feature subset that contains both discriminating features due to the use of the ReliefF ranking and unrelated features because each feature is selected from a separate grouping. This manner of feature selection should minimize both feature redundancy and the size of the feature subset used to train the base classifier.

For each fold, a J48 ensemble classifier is created from a set of trained base classifiers produced by repeating the random selection of features from each of the related groups and training using those selected features. The repeated process is done E (Line 8) times to create each of trained base classifiers. Each trained base classifier in the ensemble is applied to the test data to produce its classification result (line 11). The ensemble classifier uses a majority vote to determine the final classification result. This same process is done for each fold. The overall accuracy recorded is computed as average accuracy across all folds. For classification errors, the type of error is also recorded. Similarly, the overall sensitivity, specificity, and F-Measure are calculated as their average across all folds.

Each trained base classifier is applied to the test data for the fold to produce its classification result. The ensemble classifier then aggregates these results from the trained base classifiers using a majority vote to determine the final classification result. This same process is done for each of the folds (Line 2) and the various overall performance measures are determined as the average over all the folds. Although the classifiers that form the ensemble are created independently, each episode of learning draws from the same feature sets, with each set containing similar features. Thus, the classifiers in the ensemble are not fully independent, which is generally considered an asset for majority-vote classifiers.

IV. Experimental Results

As seen in Line 1 of the algorithm, the input parameters are S, N, G, and E where S is the fuzzy set similarity measure; N is the number of top ranked features ReliefF is to produce; G is the number of clusters in which to group related features; and E is the number of classifiers used to create the ensemble classifier. In the experiments E was fixed at 100. N ranged from 10 to 100 by 10, and G ranged from 2 to 10 by 1.

The performance of the various relatedness measures in the new feature selector is determined first using accuracy, a straightforward measure of the percentage of correct

classifications. Since in real world applications cases where types of classification mistakes are not equally undesirable, other standard measures of sensitivity, specificity, and F-measure may be used. All of these measures use a combination of the number of true positives (TP), the number of true negatives (TN), the number of false positives (FP) and the number of false negatives (FN).

Sensitivity, also referred to as recall, is the probability of correctly identifying positive samples from all the samples which are actually positive and given as

Sens (Recall) = TP / (TP + FN).

Specificity is the probability of correctly identifying the negative samples from all the samples which are actually negative and given as

Spec = TN / (TN +FP).

Accuracy is the probability of correctly identifying both positive and negative samples from all samples and given as

Acc = (TP + TN) / (TP +TN +FP + FN).

Precision is the probability of correctly identifying positive samples from all the samples which the classifier returned as positive and given as

Prec = TP / (TP + FP).

The F-measure is the harmonic mean of precision and recall. Recall is the same as sensitivity. The F-measure is given as

F-meas = 2*(Prec*Recall) / (Prec + Recall).

Figure 3 shows the performance in terms of highest overall accuracy for each measure for each of the four data sets. Table II shows the data used to create the bar chart with the N and G values producing the highest accuracy. Those shown underlined are the greatest values achieved for each data set. Note that the different performance measures may achieve their highest values at different N and G values for a similarity measure.

Breast data set: The Breast bar chart differs from the others in that Cos produces the highest accuracy at 0.816. The Cos measure, however, has the lowest accuracy for both the Leukemia and Colon data sets. IncMax follows the Cos with a 0.796 accuracy. Note also that IncMax has the lowest accuracy for the CNS data set. All *accuracies* are greater than 0.730

CNS data set: The bar chart shows the lowest accuracies are produced for the CNS data set. All measures have a greater than or equal to accuracy than 0.650 accuracy. The highest is 0.700 for the Jaccard measure and the InclusionMin. The lowest is for InclusionMax at 0.650. Three measures CCC, IncAvg and Zadeh produced the same accuracy of 0.683.

Colon data set: The bar chart follows a similar pattern to the Leukemia data set. It has the next highest accuracy with that of 0.903 for the CCC and Jaccard measures and all other measures except for Cos have the same accuracy of 0.887.

Leukemia data set: The bar chart shows the highest accuracies are for the Leukemia data set with all measures having greater than 0.900 accuracy. The highest accuracy at 0.986 occurs for all except the Cos and Zadeh measures.

Summary over data sets: The Leukemia data set has the highest accuracies for all measures followed by the Colon data set. Note that five of the seven fuzzy set similarity measures produce the same accuracy of 0.986 for the Leukemia data set. The CNS data set has the lowest accuracies for all measures.

Summary over similarity measures: The Jaccard measure has the highest or tied for the highest over three of the four data sets, i.e., the exception being for the Breast data set where the Cos has the highest accuracy of 0.816. The ties for the Jaccard occur with CCC for the Colon and Leukemia data sets at 0.903 and 0.986, respectively and with IncMin for CNS data set at 0.700. The Jaccard also ties with all three inclusion measures for the Leukemia data set at 0.986.

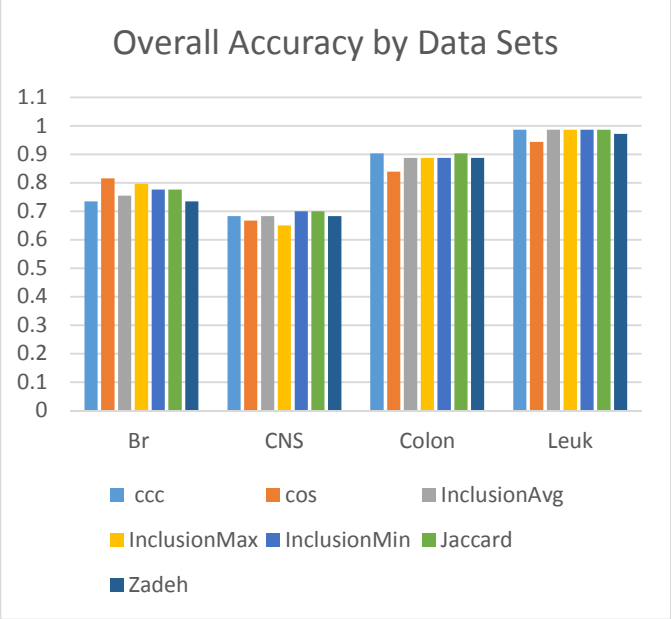

Fig. 3. Highest Overall accuracy by data sets

TABLE II OVERALL ACCURACY WITH N AND G VALUES

	Br	N	G	CNS	N	G	Col	N	G	Leu	N	G
CCC	0.735	10	5	0.683	90	9	0.903	30	4	0.986	40	3
Cos	0.816	10	7	0.667	20	9	0.839	10	4	0.944	10	2
Inc Avg	0.755	10	4	0.683	90	6	0.887	30	4	0.986	80	6
Inc Max	0.796	10	5	0.650	30	8	0.887	50	3	0.986	30	5
Inc Min	0.776	10	6	0.700	100	9	0.887	50	4	0.986	100	7
Jac	0.776	10	6	0.700	60	9	0.903	50	9	0.986	100	2
Zad	0.735	10	9	0.683	30	8	0.887	30	4	0.972	10	2

Table II shows that for the Breast data set its highest overall accuracy for each fuzzy set similarity measure occurs consistently using only the 10 highest ranked features. As can be seen in Table I, the Breast data set is balanced on the number of positive and negative cases. The Colon data set has a smaller range of N values, 10 to 50, than either the CNS or Leukemia data sets. The Leukemia data set has the greatest range of N values, 10 to 100 and yet the most tied accuracy values.

Although a maximum of G = 10 was used in the experiments, no similarity measure required that many features to achieve its highest overall accuracy. The Colon data set is very consistent across the similarity measures with the use of 4 or fewer groups except for Jaccard. The CNS data set required the greatest number of groups to achieve its highest overall accuracy for each fuzzy set similarity measure.

The bar chart in Figure 4 shows the sensitivity performance of the similarity measures for each data set. Table III provides the sensitivity values along with the N and G values where the highest sensitivity value occurred.

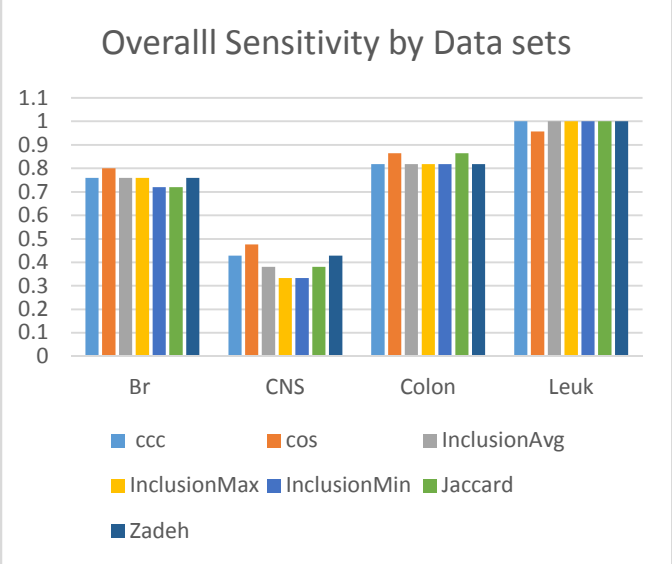

Fig. 4. Overall Sensitivity by data sets

TABLE III OVERALL SENSITIVITY WITH N AND G VALUES

	Br	N	G	CNS	N	G	Col	N	G	Leu	N	G
CCC	0.760	10	5	0.429	10	8	0.818	30	4	1.000	10	2
Cos	0.800	10	7	0.476	20	9	0.864	30	5	0.957	10	2
Inc Avg	0.760	10	4	0.381	60	10	0.818	30	3	1.000	30	2
Inc Max	0.760	10	5	0.333	10	7	0.818	30	4	1.000	30	5
Inc Min	0.720	10	4	0.333	10	6	0.818	50	4	1.000	30	4
Jac	0.720	10	5	0.381	60	9	0.864	50	9	1.000	40	2
Zad	0.760	30	8	0.429	30	8	0.818	30	2	1.000	10	2

From the bar chart it is easy to see that the smallest sensitivity values occur for all measures on the CNS data set. The greatest overall sensitivity values occur for all measures on the Leukemia data set with all measures achieving a value of 1 except for the Cos with a sensitivity of 0.957. Although the value of 1 is achieved, what varies is the N value at which it occurs. For CCC and Zadeh, it happens at N=10. For all three inclusion measures it happens at N=30 and for the Jaccard at N=40. Although Cos does not achieve a sensitivity value of 1.000 for the Leukemia data set, it has the greatest or tied for greatest overall sensitivity for the other three data sets. The InclusionMin has the lowest or ties for lowest for all data sets except for the Leukemia data set.

The bar chart in Figure 5 shows the overall specificity performance of the similarity measures for each data set. Table IV provides the specificity values along with the N and G values where the specificity value occurred. From Figure 5 it is easy to see that the smallest specificity values occur for all measures on the Breast data set. The best performing measure for the Breast data set is Cos with a 0.917 value. The worst is the CCC with a 0.792. All the others tied with a value of 0.833. The other 3 data sets have comparable specificity values, all above 0.920. The CNS data set, however, has 4 measures obtaining a specificity of 1.000, whereas the Leukemia data set has only one with specificity of 1.000 and Colon data set has none. It is interesting that the CNS dataset has the highest specificity values for all similarity measures compared to the other data sets measures except for IncMin for Colon and Leukemia data sets.

Across the fuzzy set similarity measures, again the Cos measure has the highest or tied for the highest specificity value for three of the four. Only the InclusionMin measure with a specificity of 1.000 is higher for the Leukemia data set. The CCC measure has the lowest specificity for the Breast and the CNS data sets. InclusionMax and Zadeh have the lowest for Colon. For Leukemia data set, all measures have identical specificity of 0.960 except for InclusionMin with a specificity of 1.000.

Some variation exists in the similarities' performances based on the measure most important for the evaluation or diagnosis process. With sensitivity true positives as well as how many false negatives the classifier has are considered, i.e., how often one is told they don't have a disease but that person actually does. With specificity true negatives as well as how many false positives the classifier has are considered, i.e., how often one is told they have a disease but that person actually does not. For diagnosing cancer, false negatives, or failing to diagnose cancer, are more undesirable than false positives.

The Cos similarity measure is interesting since it has the best performance with respect to sensitivity and specificity over all the data sets yet only has the highest accuracy for the Breast data set and the lowest accuracy for the Colon and Leukemia data sets. One can also see that the Leukemia data set models have the highest accuracies and the highest sensitivities across all fuzzy similarity measures. The CNS data set has the lowest accuracies and lowest sensitivities across all fuzzy similarity measures. For specificity values, however, the CNS has the highest values for six of the seven similarity measures.

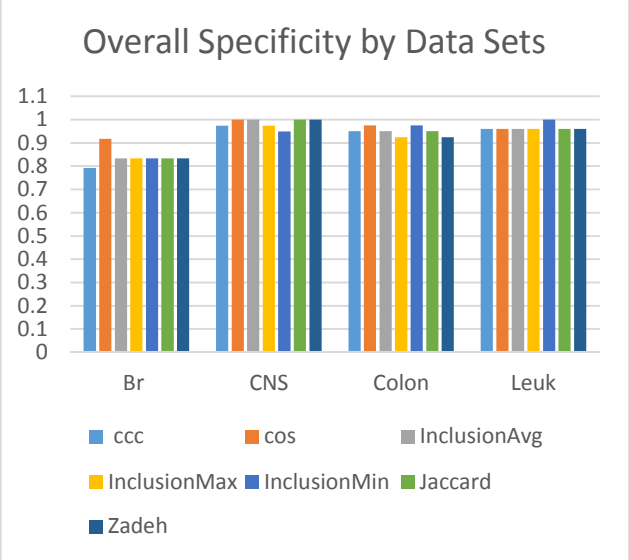

Fig. 5. Overall Specificity by data sets

TABLE IV OVERALL SPECIFITY WITH N AND G VALUES

	Br	N	G	CNS	N	G	Col	N	G	Leuk	N	G
CCC	0.792	10	7	0.974	100	2	0.950	10	4	0.960	20	3
Cos	0.917	10	5	1.000	70	2	0.975	10	2	0.960	100	9
Inc Avg	0.833	10	3	1.000	90	2	0.950	10	7	0.960	20	3
Inc Max	0.833	10	3	0.974	80	2	0.925	10	3	0.960	20	4
Inc Min	0.833	10	6	0.949	50	2	0.975	10	5	1.000	100	9
Jac	0.833	10	6	1.000	90	2	0.950	10	5	0.960	20	4
Zad	0.833	10	9	1.000	90	2	0.925	10	2	0.960	30	5

The bar chart in Fig. 6 shows the F-measure performance of the similarity measures for each data set. Table V provides the F-measure values along with the N and G values where the highest F-measure value occurred. The F-measure incorporates recall, i.e., sensitivity as accuracy does, but instead of specificity, it uses precision. It ignores the true negatives. Comparing accuracy to F-measure values, there is little difference between them over all the similarity measures for the Breast and Leukemia data sets, less than 0.015.and 0.020, respectively. For the CNS data set, however, there is a much bigger difference with accuracy having the larger values and the differences ranging from around 0.2 to 0.27. The Colon data set also has larger accuracy values with the differences around 0.05.

The larger accuracy values could be attributed to true negatives being used in its calculation where these are not included in recall and precision used in calculating the F-measure. This can also be seen in the high specificity values for the CNS data set where several values are 1.000 indicating that no false positives existed for those similarity measures.

The Cos and Jaccard combination approaches are the winners for F-measure. It could be argued they are all one needs to focus on, if this trend continues with more data sets. The Jaccard and Cos measure combination also has the highest accuracies overall the four data sets as seen in Table II.

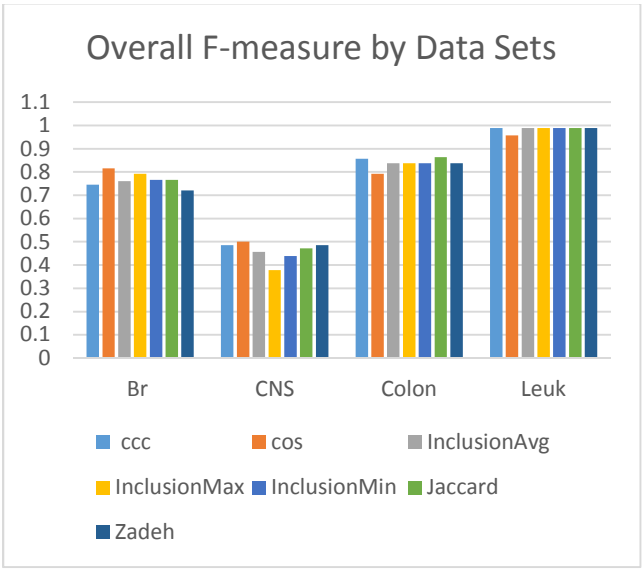

Fig. 6. Overall F-Measure by data sets

TABLE V OVERALL F-MEASURE WITH N AND G VALUES

	Br	N	G	CNS	N	G	Col	N	G	Leuk	N	G
CCC	0.745	10	5	0.486	90	9	0.857	30	4	0.989	40	3
Cos	0.816	10	7	0.5	90	9	0.792	40	7	0.957	40	2
Inc Avg	0.760	10	4	0.4570	100	8	0.837	30	4	0.989	80	6
Inc Max	0.792	10	5	0.378	10	7	0.837	50	3	0.989	30	5
Inc Min	0.766	10	6	0.438	100	10	0837	50	4	0.989	100	7
Jac	0.766	10	6	0.471	60	9	0.864	50	9	0.989	100	2
Zad	0.722	50	10	0.486	30	8	0.837	30	4	0.989	100	4

V. CONCLUSIONS AND FUTURE WORK

A study on the use of fuzzy set similarity measures along with the CCC measure to group features for machine learning has been presented. These experiments follow an approach similar to that in [7] which proposes the CCC_RSM algorithm. Two differences are the use of hierarchical clustering and a fixed number of clusters instead of using a similarity threshold based on the CCC measure as in CCC_RSM to group features.

The normalization process used in our research is done first on the training data without the use of the test sample. The test sample, however, is normalized based on the min and max of the feature values of the training data. The normalized feature values of the test sample are kept in the range [0, 1] if its feature values are outside the training data's min and max feature values.

The results show that using CCC to guide clustering is not superior to fuzzy set similarity measures. Over all the data sets and the performance measures, only for the colon data set does CCC have a higher overall accuracy than all the fuzzy set similarity measures except for its tie with Jaccard. For sensitivity, the CCC does not have the highest value for any of the data sets. For the leukemia data set, CCC does tie for high sensitivity with five other fuzzy set similarity measures with a value of 1.0. For specificity, CCC does not have the highest value for any of the data sets. For F-measure, the CCC does tie for high value of 0.989 with five other fuzzy similarity measures for the leukemia dataset.

From this analysis, fuzzy set similarity measure for the most part performs as well if not better than CCC. The results, however, do not show any single fuzzy set similarity measure to be the best over all the data sets. The Cos measure has the highest value over all the performance measures for the breast dataset. It also has the highest, or tied for the highest, sensitivity and specificity for the CNS and colon datasets and has the highest F-measure for the CNS data set. The Cos measure, however, has the lowest, or tied for the lowest values, for the leukemia data set. The Jaccard measure has the highest, or tied for the highest, accuracy for all data sets except for the breast data set and has the highest F-measure for the colon and leukemia data sets. The Inc Min measure has the highest specificity for the leukemia and colon data sets.

The best fuzzy set similarity measure is dependent on the particular data set and the performance measure being used. There is, however, a distinction in performance across the different data sets and performance measures. While it would obviously be more useful to have one measure that is best for all data sets and performance measures, the results show that the Cos measure is the most consistent high-performer over most of the data sets and many of the performance measures and that focusing exclusively on one measure such as CCC is not justified. The practitioner committed to obtaining additional levels of increased performance may find a fuzzy similarity measure that provides better performance than the CCC.

Although the CNS, Colon and Leukemia data sets are all imbalanced, no methods to address imbalance such as SMOTE [15] were used on these data sets in these experiments. SMOTE was used in [7], and their results for the CNS data set had an accuracy of 0.8864 which is much higher than any of the results reported here for CNS. Future work is to investigate how well the fuzzy set similarity measures perform after addressing class imbalance on all the data sets. Other methods of creating groups of related features, such as similarity threshold clustering as in [7] and k-means clustering, are to be used with the fuzzy set similarity measures to determine any differences in the performance of the combination of the two. More studies using different data sets are also needed to further understand the performance of fuzzy set similarity measures used in different applications of machine learning.

REFERENCES

[1] V. Boln-Canedo, N. Sanchez-Maroo, and A. Alonso-Betanzos, "Feature Selection for High Dimensional Data," Artificial Intelligence: Foundations, Theory and Algorithms. 1st edition, Spring Publishing Company, 2015.

[2] J. Önskog, E. Freyhult, M. Landfors, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning," BMC Bioinformatics, 12, 390 doi:10.1186/1471-2105-12-390, 2011.

[3] S. Wenric, and R. Shemirani, "Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies," Frontiers in Genetics, 9, pp. 297, 2018.

[4] I, Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning., vol.3, pp. 1157–1182, 2003

[5] H. Liu, L. Liu, H. Zhang, "Ensemble gene selection by grouping for microarray data classification," Journal of Biomedical informatics," 43 (1) pp, 81–87, 2010.

[6] T. K. Ho, "The random subspace method for constructing decision forests," Pattern Analysis and Machine Intelligence, IEEE Transactions on vol. 20 no.8 pp. 832–844, 1998.

[7] B. Chaudhury, D. B. Goldgof, L. O. Hall, R. A. Gatenby, R. J. Gillies, J. S. Drukteinis, "Correlation based random subspace ensembles for predicting number of axillary lymph node metastases in breast dce-mri tumors," IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2164–2169. 2015.

[8] J. M. Bland, D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," The lancet, vol. 327 (8476), pp. 307–310, 1986.

[9] V. Cross, An Analysis of Fuzzy Set Aggregators and Compatibility Measures, Ph.D. Dissertation, Computer Science and Engineering, March, Wright State University, Dayton, OH, 264 pages, 1993.

[10] P. Jaccard. "The distribution of the flora in the alpine zone", New Phytologist, vol. 11, pp. 37–50, 1912.

[11] S. Russell and P. Norwig, Artificial Intelligence: A Modern Approach 3rd Edition, Pearson Eduction Limited:Harlow, 2014.

[12] M. Robnik-Sikonja, I. Kononenko, "An adaptation of relief for attribute estimation in egression," D. H. Fisher (Ed.), Fourteenth International Conference on Machine Learning, Morgan Kaufmann, pp. 296–304, 1997.

[13] R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, John Wiley & Sons:Canada, 2001.

[14] https://www.cs.waikato.ac.nz/ml/weka/ , accessed April 2020.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research 16 2002 pp. 321–357.