Feature Space Enrichment by Incorporation of Implicit Features for Effective Classification

Abhishek Srivastava
Dept. of Computing Science
University of Alberta
Edmonton, Canada
sr16@cs.ualberta.ca

Osmar. R. Zaïane
Dept. of Computing Science
University of Alberta
Edmonton, Canada
zaiane@cs.ualberta.ca

Maria-Luiza Antonie
Dept. of Computing Science
University of Alberta
Edmonton, Canada
luiza@cs.ualberta.ca

Abstract

Feature Space Conversion for classifiers is the process by which the data that is to be fed into the classifier is transformed from one form to another. The motivation behind doing this is to enhance the "discriminative power" of the data together with preserving its "information content". In this paper, a new method of feature space conversion is explored, wherein "enrichment" of the feature space is carried out by the augmentation of the existing features with new "implicit" features. The modus operandi involves generation of association rules in one case and closed frequent patterns in another and the extraction of the new features from these.

This new feature space is first made use of independently to feed the classifier and then it is used in unison with the original feature space. The effectiveness of these methods is subsequently verified experimentally and expressed in terms of the classification accuracy achieved by the classifier.

1. Introduction

Classifiers are computational models that have the ability to predict the class of a data item on the basis of the values of its characteristic attributes [2]. For doing this, the classifier has to be first "trained" with a set of representative data items. Once trained, the classifier has the ability to assign the most suitable class to an unlabelled data item. The accuracy of a classifier is measured by "hiding" the class labels of a set of labeled data items and monitoring the percentage of these data items that the classifier correctly classifies.

The classifier that has been used in our work is the Support Vector Machine (SVM) [3] for its reputation to be one of the best, if not the best, classifier in many real applications. The SVM separates data into classes by attempting to find a linear "maximum margin hyper-plane". If the data is linearly separable, such a hyper-plane is found and the data is classified. If the data is not separable linearly then the SVM "raises" the data to a much higher dimension making use of special functions called kernels. At a higher dimension when the data becomes separable the SVM finds the suitable hyper-plane.

The "feature space" of a classifier refers to the attributes of the data item that are made use of by the classifier for distinguishing one item from another. For example, a set of creditors at a bank may be distinguished from one another on the basis of their income, age and education level. In this case, the feature space comprises : {Income, Age, Diploma}.
Feature space conversion therefore implies a *change* in the set of attributes that are thus used. This *change* may be brought about by modifying the current attributes used, by making use of an entirely new set of attributes, or a combination of the two.

Normally, the feature space that is made use of comprises the *explicit* features *i.e.* the original characteristic attributes of the data item. This paper makes an attempt to *enrich* this feature space by incorporating certain *implicit* features which are not obvious but which have to be extracted from the available attributes. We explore two possible methods of doing this: (1) a method based on generation of association rules, and (2) one based on closed frequent patterns. The first of these entails the generation of association rules which are relationships that exist between different data items in a transactional database such that the presence of one item implies the other or the presence of a combination of items implies the presence of a third item [12]. For example, weather: sunny and day: Saturday implies mood: happy.
As an association rule, this is written as:

Sunny ^ Saturday → happy
(antecedent) (consequent)

The *support* of an association rule X → Y refers to the fraction of transactions that contain (X U Y) items and the *confidence* of X → Y is the fraction of the transactions containing X that also contain Y.

Note that in this paper we are not claiming a new associative classifier [6, 14, 15] (*i.e.* a classifier based on association rules) but investigate feature space enrichment to potentially improve any classifier; in our case we use SVM.

In this paper, each data point which comprises a set of attributes and a class label, is considered a transaction and the association rules generated are those between the union of the different attribute values and the class labels. Having been generated, the rules are "filtered" to obtain only those that have the class-labels as the consequent *i.e* . the implied value. From these select association rules, two sets of features are derived: Rule Based features and Class Based features [4]. These will be discussed in detail later.
This converted feature space is subsequently made use of to train the classifier, in this case SVM, and the accuracy of the classifier is monitored.
The rule based and class based features are first used to train the SVM independently. They are further used in combination with the original feature space of the data. The variation in accuracy of the classifier is studied over different values of minimum support threshold (*i.e.* by using features generated from association rules whose support is above the minimum support threshold).

In addition to this, closed frequent patterns are also made use of, to generate new features [13]. A group of items X in a transactional database is a closed frequent pattern if X occurs in the database more frequently than the minimum support threshold and there is no *proper* super-set of X that has the same support as X.

All the frequent patterns for the concerned dataset are generated and from these patterns, a new feature vector is fabricated for each original data vector. The new feature space thus obtained, like in the case of the association rules ones, is first used independently to train the SVM and then in combination with the original features.
Some work on feature space augmentation or enrichment has been done but limited and not necessarily related to our focus of study. For instance feature space augmentation was investigated in the context of classification with taxonomies [16]. The authors of [17] investigating image clustering highlight the need for feature space augmentation in the context of image datasets but do not exploit the possibility.

Relatively less work has been done on feature space augmentation of the kind we are dealing with. Rather most of the related work concentrates on the "trimming" of the feature space so as to effectively handle large volumes of high dimensional data. It is referred to as feature selection. This has especially been done in text categorization. Yang *et al.* reduce the dimensionality of the feature space of text documents by quantitatively expressing the relevance of terms using Information Gain and the χ^2-test methods and expressing the content using the highly relevant terms only [9]. Koller *et al* . in their work attempt to transform the feature space of text documents by first creating a hierarchy of topics and then merging sufficiently *close* topics to each other [10]. From this reduced number of modules, representative terms are chosen as features. Scott *et al.* explore the "syntactic and semantic" relationships that exist between words in a text, rather than the morphological relationship as was normally done [11]. All synonymous terms were mapped to one feature. This way, they were able to substantially reduce the feature space. One recent work that does concentrate on feature space augmentation is that of Cheng *et al.*[8]. They map a relationship between minimum support threshold and information gain, and modify the original feature space by generating closed frequent patterns corresponding to the optimal support threshold and combining them with the original features.

2. Feature Space Conversion

As mentioned, two broad methodologies of feature conversion are made use of in this paper. The new features are mainly used to enrich the feature space *i.e.* they are used in combination with the original features although we also briefly analyze their respective independent influence on the classifier accuracy. The two categories of feature conversion being dealt with here are:

- Association rules based features.
- Closed frequent patterns based features.

2.1. Association Rules based Features

The methodology followed to carry out feature space conversion involves first the generation of association rules from the data-set, followed by filtering out irrelevant rules, and finally the construction of the rule-based and class-based features.

2.1.1. Generation of the Relevant Association Rules. The data-set that is made use of for classification usually consists of a set of data points each represented by a unique *vector*. This vector comprises the attribute values of the data-point as also the class to which the data point belongs. This is illustrated in the following:

$$X: x_1, x_2, x_3, \ldots. x_n \quad C_x$$

x_i (i = 1-n) represents the values of n attributes of data point X and C_x represents the class to which X belongs.

The association rules are generated by considering each data vector to be a transaction, and the attributes and class labels as the data items. Let us consider a simple example of a very small data-set:

2, 12, 1, 67, 3, 6, 7, 23, 9, 8, C_1
54, 7, 8, 22, 1, 9, 78, 12, 91, C_1
7, 1, 89, 4, 22, 12, 3, 9, 54, 2, C_2
1, 123, 7, 8, 3, 35, 65, 2, 9, 66, C_3
2, 1, 4, 6, 89, 3, 56, 3, 88, 9, C_2
7, 12, 95, 16, 9, 1, 56, 78, 70, C_1

The attributes C_i are the class labels.

From this data-set, the association rules are generated. Further, from these rules the only rules that are relevant are the ones that have a class label as the consequent. All other rules are "filtered" out. Below is a simple example of a possible set of relevant rules.

Table 1. Example of a set of relevant rules

$9 \wedge 78 \wedge 12 \rightarrow C_1$	Sup. = 33.33%, Conf. = 66.67%
$7 \wedge 1 \wedge 9 \rightarrow C_1$	Sup. = 50%, Conf. = 60%
$3 \wedge 89 \rightarrow C_2$	Sup. = 33.33%, Conf. = 100%

The more practical approach however is to make use of tools that directly generate association rules with the constraint that the consequent should be a class label rather than generating all the association rules and filtering out the irrelevant ones.

2.1.2. Construction of the New Feature Space. This portion of the methodology is the *crux*. Based on the selective association rules obtained in the previous two steps, two new feature spaces are generated:

(i) Rule Based feature Space.
(ii) Class Based Feature Space.

2.1.3. Rule Based Features. In the rule based feature conversion method, for every data point in the data set, a new feature vector is generated. The steps followed in doing this are simple. All the relevant association rules obtained in the previous steps are scanned. The rules whose antecedent is *contained* in the original attributes of the data point are checked. In the new feature vector every relevant rule generated is assigned two fields. The first field takes a value of 1 if the rule is marked (*i.e.* its antecedent is contained in the attributes of the data point) otherwise it takes a value 0. The other field is assigned the confidence value of the respective rule.
A simple example follows: Let us consider the set of relevant rules in Table 1.

Let the data points be :

23, 45, 8, 1, 3, 54, 7, 123, 89, 9, 17, C_3
1, 4, 3, 12, 78, 9, 7, 52, 654, 89, 90, C_2

The rules whose antecedents are contained in the attributes of the first data point are the second and the third rule in Table 1 whereas the antecedents of all 3 rules are contained in the second data point.

The rule based feature vectors for these data points thus become:

Rule 1		Rule 2		Rule 3	
Conf.	Marked	Conf.	Marked	Conf.	Marked
66.67	0	60	1	100	1

Rule 1		Rule 2		Rule 3	
Conf.	Marked	Conf.	Marked	Conf.	Marked
66.67	1	60	1	100	1

2.1.4. Class Based Features. The methodology for obtaining the class based feature vector is similar to the rule based one. Here, too the rules are scanned and those rules whose antecedents are contained in the data point are marked.

The feature vector in this case, however, has a different construction. All the rules that are marked are examined for their respective consequents (*i.e.* class labels in this case). The new feature vector constructed, has two fields allocated for each *unique* class that the marked rules have as their consequents. The first field is given the value of the number of rules that have that class label as their consequents, and the other field is allocated the average confidence values of all the rules to which that class applies.
Continuing with our example :
Two rules are contained in the first data point, the second rule and the third rule. Each of these rules has a unique class label in its consequent, rule 2 has class C_1, and rule 3 has class label C_2. Therefore there would be four fields in the new feature vector, two corresponding to class C_1 and two to C_2. Whereas in the second data point, all three rules are contained. The first two rules have the same class label C_1, therefore the "Avg. Conf." field in this case becomes 63.33 (average of 66.67% and 60% of the first and second rules respectively) . The third rule is the only rule with the label C_2 and therefore the Average Confidence is equal to the confidence of the rule.

The feature vectors obtained are shown below :

. Class 1 Class 2

No of rules	Avg. Conf.	No of rules	Avg. Conf,
1	60	1	100

Class 1 Class 2

No of rules	Avg. Conf.	No of rules	Avg. Conf,
2	63.33	1	100

2.2 Closed Frequent Patterns based Features

This approach of generating features is more straightforward. Here, the first step is the generation of all the closed frequent patterns corresponding to the set minimum support threshold. The patterns generated are then numbered from 1 to n (n being the number of closed frequent patterns). Next, given a data vector, all the generated patterns are scanned to see if they exist within the data vector. If a certain pattern does exist within the data point, its corresponding number is assigned a "1" in the new feature vector, else it is assigned a "0".

Again, considering a simple example. Let the set of closed frequent patterns generated that are subsequently numbered be:

1.	7,9
2.	23,48,2
3.	1,7,17
4.	53,49,78,61

Let the original data vector be:

23, 45, 8, 1, 3, 54, 7, 123, 89, 9, 17, C_3

The patterns numbered 1, and 3 exist in the data vector whereas pattern number 2, and 4 are absent. Therefore the new feature vector becomes as shown below. Here the first field corresponds to the first pattern. Since it does exist within the data vector, the field is assigned a "1". The other fields similarly correspond to the numbered patterns and are assigned relevant values.

1	0	1	0

3. Experiments

The experiments were conducted on 4 datasets: Pima, Glass, Hayes-Roth, and Lymphography obtained from the UCI Machine Learning Repository [5]. The association rules corresponding to the closed frequent patterns for these datasets were generated using an 'in-house' software. Several groups of association rules were generated corresponding to different minimum support thresholds within a restricted range (15 – 25% for Glass dataset, 1-51% for Lymphography, and 1-11% for the others). The range was restricted because beyond the upper-bound, very few rules were generated to be of relevance. The same software was also made use of to generate the closed frequent patterns. The minimum support threshold was deliberately kept the same as that for the association rules.

The rule based and class based feature conversion was carried out using the association rules generated from the closed frequent patterns only and the closed pattern based feature conversion was carried out using the closed frequent patterns generated.

The classifier made use of was LIBSVM [7]. 10-fold cross validation (which is performed by the classifier itself) was considered and the accuracy was calculated on the training data. The kernel used was the linear kernel.

All the new feature spaces obtained were first used independently to train the classifier and then combined with the original features and the combined features were used to train the classifier. Subsequently, the following combinations of features were also used:

1. Rule-based + Class-based
2. Rule-based + Class-based + Original Features
3. Closed Frequent Patterns based features + Rule-based + Class-based + Original features (*i.e.* all features)

The classification accuracy achieved is plotted against the different minimum support thresholds used for the 4 datasets.

3.1 Experimental Results

The results are plotted as the classification accuracy achieved as a percentage against the minimum support threshold also expressed as a percentage. The plots in Figures 1, 2, 3, and 4 are the accuracy versus support threshold for the four datasets respectively. The accuracy achieved using all the combinations mentioned above are plotted *i.e.* the new features independently, the new features in combination with the original features, and the three combinations indicated explicitly above. The accuracy achieved by the original features is also indicated in each of the graphs as a reference by
a straight dotted line (above the line indicates an improvement in accuracy)

Figure1 shows the plot for the Glass dataset. Here, a demarcation is clear between the new features in

combination with the original (which return a much higher accuracy) and the new features taken alone. Figures 2 and 3 show the plots for the Hayes-Roth, and the Pima datasets respectively. In these cases the difference between the two groups is not as marked as in Glass but broadly, the combined (*i.e.* the new features + original) perform better than the new features alone. Figure 4, for the lymphography dataset is a contrast, with the new features taken alone returning better accuracy results than when taken in combination with the original features. There is therefore no clear 'winner'. The only point that is consistent is the fact that the augmented features usually perform better.

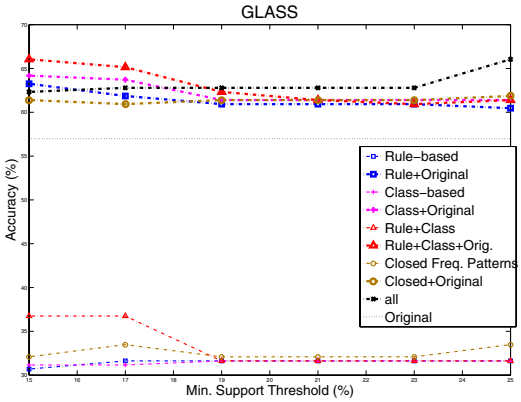

Figure 1. Accuracy versus min. support threshold for the Glass dataset

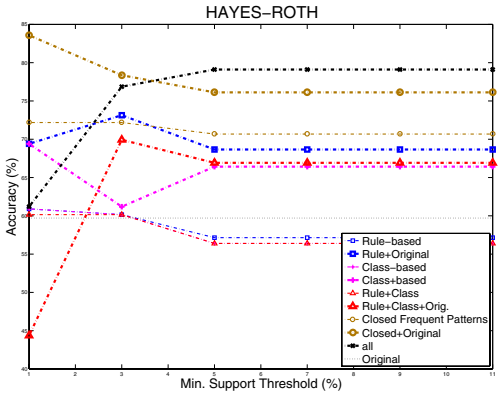

Figure 2. Accuracy versus min. support threshold for the Hayes-Roth dataset

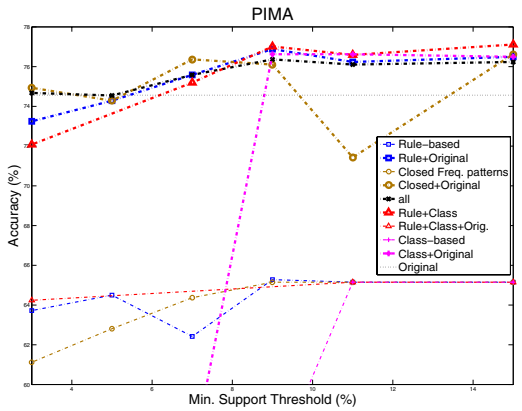

Figure 3. Accuracy versus min. support threshold for the Pima dataset

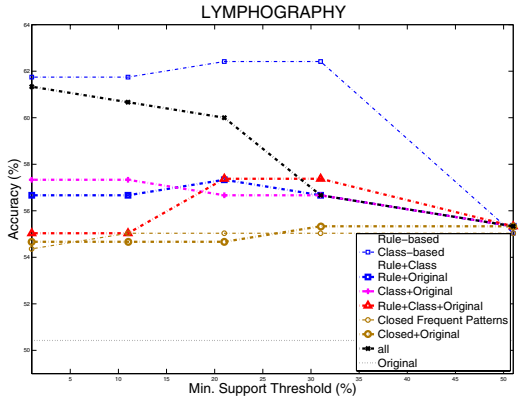

Figure 4. Accuracy versus min. support threshold for the Lymphography dataset

Studying the behaviour of the accuracy achieved by the original features augmented by the various feature spaces, reveals certain notable points. The average (over the four data-sets) variation of the (rule-based + original) feature space is interestingly almost identical to the average variation of the original feature space augmented by the closed frequent patterns based features. The other feature spaces when combined with the original features have a tendency to gradually improve the accuracy achieved, with increasing minimum support threshold. This is illustrated in Figures 5, 6, 7, 8, 9. This behaviour of the feature spaces would compel us to conclude that the best accuracy results may be obtained using the features generated keeping the minimum support threshold as high as practicably possible, and using this feature space in combination with the original features.

Our approach is found to better the accuracy achieved by the original feature space in all the data-sets. It is to be however noted that most of the time the

improvement in accuracy is achieved only when new features generated enrich the original feature space. Independently each of the feature spaces falls short of the original features.

Figure 10 compares the best accuracy results obtained by each of the feature spaces in different combinations as also the original feature space.

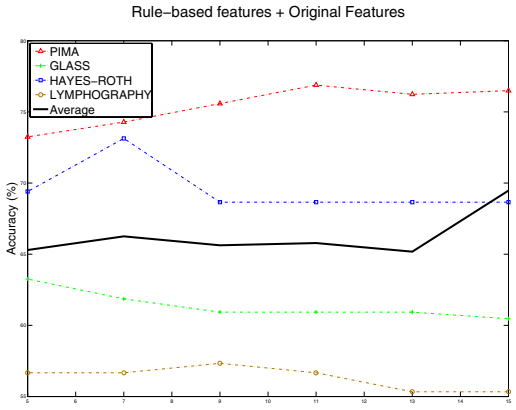

Figure 5. Behaviour of the Original features when augmented with the Rule-based features.

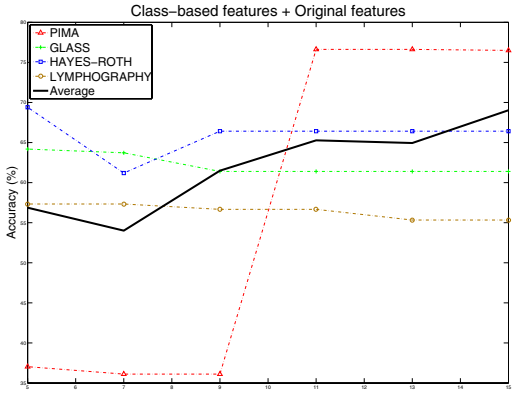

Figure 6. Behaviour of the Original features when augmented with the Class-based features.

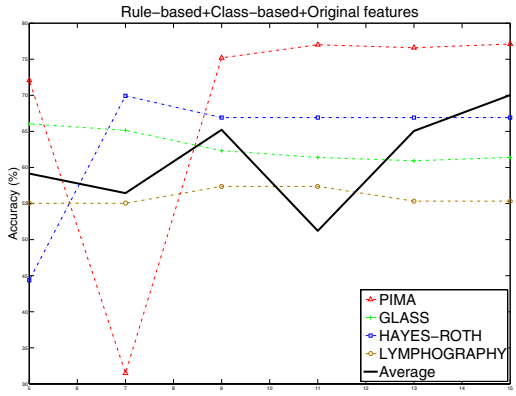

Figure 7. Behaviour of the Original features when augmented together with the Class-based and the Rule-based features.

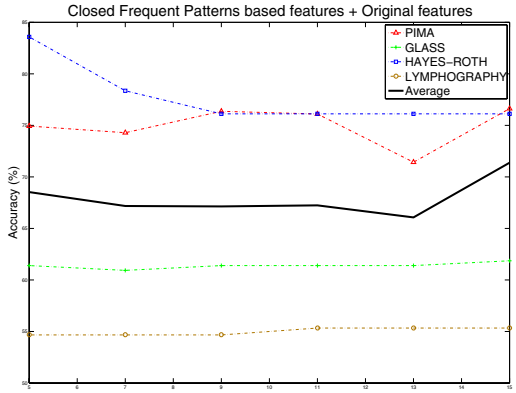

Figure 8. Behaviour of the Original features when augmented with the Frequent Closed Patterns based feature space.

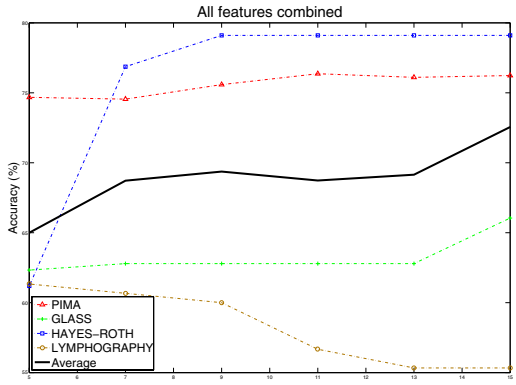

Figure 9. Behaviour of the Original features when augmented with all the new Feature Spaces together.

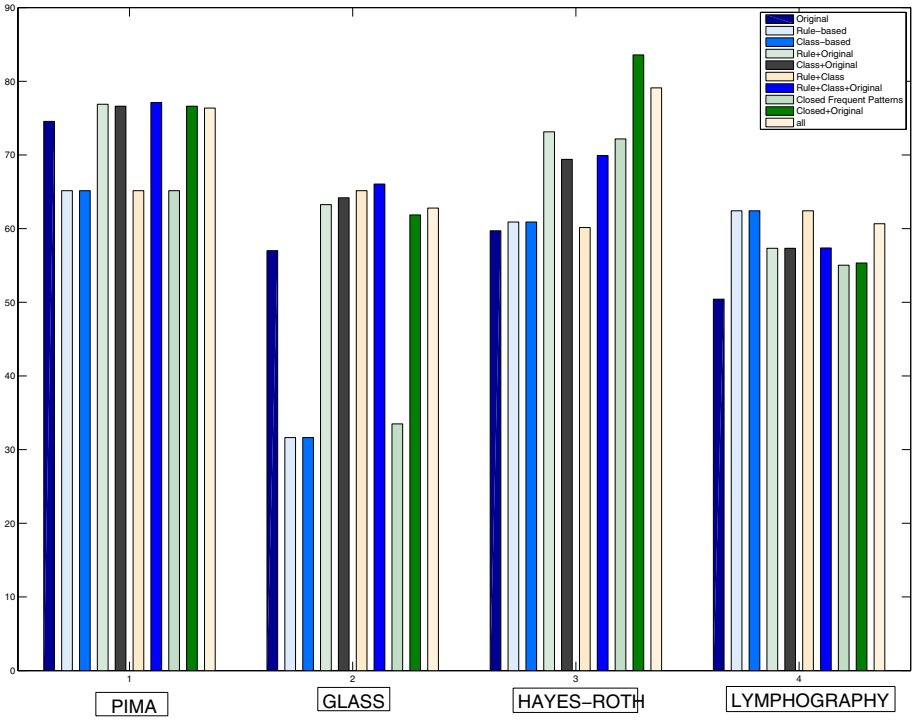

Figure 10. A comparative chart of the results of different feature spaces (all values in %)

Table 2: Accuracy values for different combinations of feature spaces

	O	R	C	R+O	C+O	R+C	R+C+O	Cl	Cl+O	Cl+O+R+C
Pima	74.56	65.15	65.15	76.88	76.62	65.15	77.11	65.15	76.62	76.36
Glass	57.00	31.63	31.63	63.26	64.19	65.15	66.05	33.49	61.86	62.79
Hayes.	59.70	60.90	60.90	73.13	69.40	60.15	69.92	72.18	83.58	79.10
Lymph.	50.42	62.42	62.42	57.33	67.33	62.42	57.38	55.03	55.33	60.67

R: Rule –based; **C**: Class-based; **O**: Original ; **Cl**: Closed Freq. Pat. based

5. Conclusion

A new approach to feature space conversion for classifiers was proposed and subsequently tested on a few datasets in this paper. It was found to substantially improve upon the accuracy achieved by the original feature space alone. In spite of this we were unable to determine one clear 'clear' amongst the methodologies, and were able to more broadly conclude that the enrichment of the original feature space with the new features returned better results than considering either of the features separately.

More important however was the fact that this paper has opened doors to a fresh approach to feature conversion using association rules and frequent patterns. Future work could concentrate on the

incorporation of further "implicit information" from these rules and patterns into the feature vector.

Studying the behaviour of the classification accuracy achieved by using the various transformed feature spaces, against minimum support threshold , we were able to conclude that the best results would be returned using the features generated from a small number of highly frequent association rules or patterns in combination with the original features.

6. References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In *Proc. of SIGMOD*, 1993, pp. 207-216.

[2] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", *Morgan Kaufmann*, 2001.

[3] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[4] M. –L. Antonie, O. R. Zaiane, and R. C. Holte, "Learning to use a learned model: A two-stage approach to classification", The Sixth IEEE *International Conference on Data Mining (ICDM'06)*, 2006.

[5] C. Blake and C. Merz, "UCI repository of machine learning databases", http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[6] Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, Application of Data Mining Techniques for Medical Image Classification, in *Proc. of Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001)* in conjunction with *Seventh ACM SIGKDD*, pp. 94-101, San Francisco, CA, August 26, 2001

[7] C. C. Chang, and C. –J. Lin, "LIBSVM: a library of support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[8] H. Cheng, X. Yan, J. Han, C. –W. Hsu, "Discriminative frequent pattern analysis for effective classification", In *Proc. of ICDE*, 2007.

[9] Y. Yang, and J. O. Pedersen, "A comparative study on feature selection in text categorization", In *Proc. of ICML*, 1997.

[10] D. Koller, and M. Sahami, "Hierarchically classifying documents using very few words", In *Proc. of ICML*, 1997.

[11] S. Scott, and S. Matwin, "Feature engineering for text classification", In *Proc. of ICML*, 1999.

[12] Agrawal R, Srikant R. "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99.

[13] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, "Discovering frequent closed itemsets for association rules", In *Proc. of ICDT*, 1999.

[14] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98*) New York, USA, 1998.

[15] W. Li, J. Han, J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules" , In *Proc. of IEEE-ICDM*, 2001.

[16] Flavian Vasile, Adrian Silvescu, Dae-Ki Kang, Vasant Honavar, "TRIPPER: Rule learning using taxonomies", In *Proc. of Tenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06)*, pp. 55-59, Singapore, April 9-12, 2006

[17] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, Unsupervised Image-Set Clustering Using an Information Theoretic Framewor, *IEEE Transactions on Image Processing*, pp 449-458, February, 2006