Mining Breast Cancer Data with XCS

Faten Kharbat
Computing Department
Zarqa Private University
Zarqa, Jordan
Faten.West@gmail.com

Larry Bull
School of Computer Science
University of the West of England
Bristol BS16 1QY, U.K.
+44 (0)117 3283161
Larry.Bull@uwe.ac.uk

Mohammed Odeh
School of Computer Science
University of the West of England
Bristol BS16 1QY, U.K.
Mohammed.Odeh@uwe.ac.uk

ABSTRACT
In this paper, we describe the use of a modern learning classifier system to a data mining task. In particular, in collaboration with a medical specialist, we apply XCS to a primary breast cancer data set. Our results indicate more effective knowledge discovery than with C4.5.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search – backtracking, control theory, dynamic programming, graph and tree search strategies, heuristic methods, plan execution formation and execution, scheduling.

General Terms
Algorithms, Measurement, Performance, Experimentation.

Keywords
Classification, Genetic Algorithm, Learning Classifier System, Medical Informatics.

1. INTRODUCTION
Learning Classifier Systems (LCS) [8] have been successfully used for data mining within a number of different medical domains, beginning with Bonelli and Parodi's [5] work using their 'Newboole' system. More recently, Wilson's XCS [29] was favorably compared to a number of other popular machine learning algorithms over numerous well-known benchmark data sets [3][27]. XCS has been applied over the three Wisconsin Breast Cancer Datasets [4] and again achieved competitive results [1]. Other examples include a Newboole-like system, termed EpiCS [10], which was found to classify novel cases in another medical domain more effectively than decision rules derived from logistic regression [9].

This paper presents a complete knowledge discovery process based on XCS using a breast cancer dataset obtained from a UK health trust. This structure includes exploiting, compacting, and evaluating the generated knowledge. It outlines the applicability of using XCS in a medical decision support task aimed at improving the diagnosis of primary breast cancer.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

2. THE DATASET
Although breast cancer research has been developing recently, the challenge has been to shift from gathering data to finding hidden patterns and trends that are most relevant to cancer diagnosis. Primary breast caner diagnosis is a major challenge to oncologists who treat breast cancer since it is the first stage from where the cancer develops. Primary breast cancer refers to cancer that has not yet spread outside the breast. The Frenchay Hospital in Bristol, in the United Kingdom started building a database for primary breast cancer in 1999. Since then they have been developing their research studies and improving their treatment based on their outcomes and results. Thus, this investigation exploits one of their datasets which is complex and useful, as Frenchay hospital has been using it and relying on it since it contains accurate collected pathological data related to their patients.

The Frenchay Breast Cancer (FBC) dataset is a real-domain dataset which has a description of pathological data for women with primary breast cancer. The development of the FBC dataset started in 2002 by Dr. Mike Shere, a consultant breast specialist in Frenchay Hospital, Bristol, UK.

Every case in the FBC is represented by 45 binary and the categorical attributes, which collectively describe the status of breast cancer in a certain patient. For the purpose of this investigation, 1150 cases from the FBC dataset were used in this knowledge discovery process. The diagnosis for each case is the cell grade, which determines the aggressiveness of the breast cancer stage and has the three grades G1, G2, or G3, with the distribution of 15.2%, 48.3%, and 36.5%, respectively as shown in Figure 1. The cell-grade is usually calculated by summing up some histological characteristics of breast carcinoma; and more specifically, it is the sum of the following attribute's values: Tubule-Formation-Score, Nuclear-Pleomorphism-Score, and Mitotic-Figure-Score.

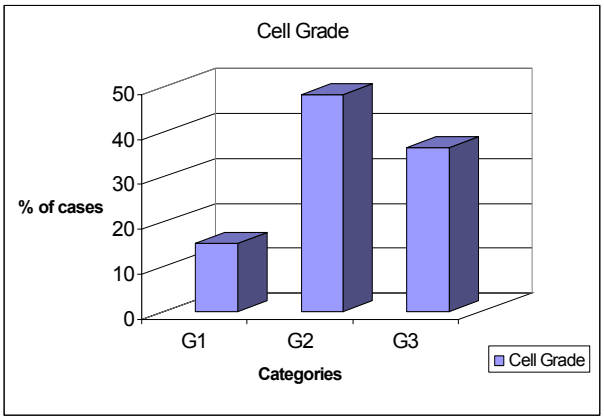

Figure 1. Distribution of Classes in FBC Dataset.

3. DATA PREPERATION

3.1 Data Pre-Processing

Data pre-processing takes place before applying machine learning techniques to solve some limitations and barriers found within the original data. This process should transform the original data into a more useful form [7]. In general, the problems within the original data extend from the existence of irrelevant attributes to the existence of multi-level noise, which prevents the knowledge discovery process from being successful.

Data pre-processing techniques vary, and there exist a huge number of techniques with different algorithms, as each rectifies a specific problem and suits certain machine learning techniques. Obviously, each algorithm has its strength and weakness points that may affect the original dataset. Filtering, noise modelling, feature selection, and data fusion are some of these techniques.

However, data pre-processing is a time consuming task given the need to (e.g., [7]): (1) determine what problems occur in the selected data, (2) determine the needed pre-processing techniques and select the best suitable algorithms for the used machine learning technique, (3) apply these over the original dataset to arrive at better resultant dataset. As this is not a closed or bounded problem, pre-processing the data is not treated here. This is because the aim of this investigation is to assess, evaluate, and compare the ability of XCS, along with other learning techniques, to classify and deal with raw data. Moreover, testing and evaluating each pre-processing technique with different learning techniques including LCS is beyond the scope of this research.

In the following sections, a simple preparation procedure is applied to setup the data in a suitable form. Formatting, decoding, and solving the imbalance problem within the FBC dataset were carried out in this phase.

3.2 Data Formatting and Decoding

Three types of attributes are used within the FBC dataset and these are: numeric, boolean and categorical. As in the related literature (e.g., [26]), numeric attributes are normalised between the values of 0 and 1, where a value X is decoded into X' using the minimum and maximum values of the attribute interval as follows:

$$X' = \frac{X - \min Val}{\max Val - \min Val} \quad (1)$$

The values in categorical and boolean attributes are decoded into values {0, 1, 2… *n-1*,#} where *n* is the number of the possible values of the attribute. Table 1 shows the decoding with a simple example of each attribute type.

Table 1. Attribute types: decoding and examples.

<table>
<tr><th>Attribute Type</th><th>No of Attributes</th><th>Decoding</th><th>Example</th></tr>
<tr><td>Numeric</td><td>15</td><td>Normalise values between the of 0 and 1</td><td>Attribute: Age
<table><tr><td></td><td>min</td><td>max</td></tr><tr><td>Original values</td><td>21</td><td>98</td></tr><tr><td>Decoding values</td><td>0</td><td>1</td></tr></table>Example: 55→ 0.44</td></tr>
<tr><td>Boolean</td><td>13</td><td>Ternary</td><td>Attribute: DCIS-Necrosis
<table><tr><td>Original values</td><td>true</td><td>false</td><td>Any</td></tr><tr><td>Decoding values</td><td>0</td><td>1</td><td>#</td></tr></table></td></tr>
<tr><td>Categorical</td><td>17</td><td>0, 1, 2… n-1,#</td><td>Attribute: Core-Biopsy-B-code
<table><tr><td>Values</td><td>B3</td><td>B4</td><td>B4b</td><td>Any</td></tr><tr><td>decoding</td><td>0</td><td>1</td><td>2</td><td>#</td></tr></table></td></tr>
</table>

3.3 The Imbalance Problem

The class imbalance problem is a well-known problem, which occurs when the classes' representation is unequal in the dataset, and thus the classes' frequency is significantly unbalanced. The dominant class, which is represented more frequently within the dataset, is referred to as the majority class. Other classes which are represented by smaller sets are referred to as the minority classes. It is believed that this problem may hinder most of the learning algorithms from achieving high accuracy performance [2].

However, this problem seems to have relations with other real dataset problems. For example, [32] revealed that class imbalance

is not the only factor responsible for holding back the classification performance, but also affecting the degree of overlap between classes. Thus, solving the imbalance problem will not always increase the learning algorithm performance as there are other problems to be considered. Japkowicz [17][13] has argued the same in that the imbalance problem may not be the main problem. But they focused on the problem of the existence of small disjuncts which correctly cover few data elements. These small disjuncts commonly are prone to higher error than large disjuncts. Several approaches were suggested to solve such problems in [13].

Although such studies have taken place, the other suggested problems (i.e., small disjuncts, etc.) do not have a clear simple solution, especially for a real-domain problem [13]. Therefore, dealing with them could just complicate, transform, or interpose the original dataset. Moreover, the imbalance problem still influences the performance of learning systems considerably; and therefore, this investigation will treat this problem as it may cause difficulties to learn concepts related to the minority classes.

One category of supervised techniques that has been broadly used to address the class imbalance problem is related to the methods in which fractions of the minority and majority data elements are controlled via under-sampling and/or over-sampling, so that the desired class distribution is obtained in the training set [15]. Using the under-sampling technique, the data elements associated with the majority class are reduced [14]; and therefore, the size of the dataset is reduced significantly. Alternatively, over-sampling is usually applied to increase the number of data elements of the minority classes [14], which will result in increasing the size of the dataset.

It was suggested in [33] that using over-sampling is required if the (majority/minority) ratio is very high. Moreover, Batista et al. [2][32] suggested that the over-sampling method may generate more accurate results than the under-sampling one. More precisely, random over-sampling showed competitive results than those more complex ones within their experiments. Recent results with XCS confirm the effectiveness of the simple over-sampling technique over more complicated ones [31].

For the FBC dataset, it can be seen that the class G2 is the major class and G1 is the most minor one with the ratio between G2 class and G1 (G2/G1) is 3.17 and the ration between G2 class and G3 (G2/G3) is 1.37; hence, balancing this dataset is required. After [31], the random over-sampling technique is chosen where random cases from the minority classes are selected and replicated so that all the classes in the dataset are represented by the same ratio.

3.4 Missing Data

Missing data is another problem that occurs within real datasets. Within this research, missing data is treated while learning by XCS. In [12] it is reported that XCS is stable across all the missing value types and densities which were illustrated in [11], and therefore this investigation does not analyze the density of the existing missing values or their types within the FBC dataset. However, it can be seen that there are different types of missing values within the dataset which vary from simple uncollected elements to more complex missing values of the attributes. For example, the *specimen type* which has only five missing values is an example of simple missing values which may be attributed to unavailable data or just an error in data entry. This was referred to in [11] as missing completely at random. Another example is the *DSIC,* a dependant attribute that is *Histology*-based. If the *Histology* has the value of *M85203*, then the value of *DSIC* will not exist, else it will have a value. This is not a missing data that is not collected or corrupted in some way; but it is the nature of the attributes which have an entire dependency. However, in this investigation this is also considered as missing data.

Missing values are treated using the Wild-to-Wild mechanism [12]. While creating the match set [M], the matching process is performed in which the missing values in an input are handled by replacing them with don't care symbols or general intervals, and therefore they match any value. Covering an attribute is handled by assuming any missing attribute is a "don't care" or the most general interval for the attribute.

4. CLASSIFICATION

4.1 Well-known Techniques

Experiments were performed using the well-known and traditional classification techniques namely, Bayesian Network Classifier [16], Sequential Minimal Optimization [34], and C4.5 [24] using the Weka software [30]. These techniques were chosen for their performances over Frenchay dataset and because they are widely used within the machine learning community. All experiments were performed using the default parameters setting in Weka with the ten-fold cross validation [20].

The Bayesian Network Classifier is a well-known supervised learning technique and is one of the probabilistic, graph-based approaches to reasoning under uncertainty. It has shown competitive results in different cancer application domains such as the prediction of survival of patients with malignant skin melanoma [25], and the identification of 33 breast cancer risks [23]. In [22] the Bayesian Network was found to perform comparatively better than Neural Networks and logistic regression models in addition to its ability to explain the causal relationships among the variables. In the case of the FBC dataset, the accuracy performance for the Bayesian Network Classifier is 70.38%± 5.15. It can be noted that the Bayesian Classifier was tested and found lacking compared to the Bayesian Network Classifier; the accuracy performance for the Bayesian Classifier is 63.93%± 4.64.

The Sequential Minimal Optimization (SMO) is an optimization algorithm that quickly solves the Support Vector Machine (SVM) quadratic programming (QP) problem without any extra matrix storage and also without invoking an iterative numerical routine for each sub-problem [16]. Also, it has been used successfully in lung cancer to aid diagnosis [21], in addition to its competitive results for breast cancer diagnosis. After applying SMO over the FBC dataset, the performance was found to be 72.50%± 3.82.

C4.5 is a well-known decision tree induction learning technique which has been used heavily in the machine learning and data mining communities. The output of the algorithm is a decision tree, which can be represented as a set of symbolic rules. For the current dataset, results showed that the C4.5 technique achieved 77.4%±3.33 classification accuracy with an average tree size of 101.51± 21.95 and 70.51± 15.9 as the number of obtained rules.

Table 2 shows a summary of the accuracy performance for the above three classification techniques. It can be seen that C4.5 achieved the highest performance among the selected classification techniques. Therefore, the generated rules (knowledge) from C4.5 are to be evaluated and compared with the rules obtained using XCS. This has been performed using the domain specialist to critically report on the results achieved as explained later.

Table 2. Accuracy performance for benchmark classifiers.

Technique	Accuracy	Tree size	No. rules
Bayesian	70.38%± 5.15	-	-
SMO	72.50%± 3.82	-	-
C4.5	77.4%± 3.33	101.51± 21.95	70.51± 15.9

4.2 XCS

This section shows the behaviour of XCS as described in [6] in learning the FBC dataset. Since the attributes in the dataset are divided into three data types: binary, categorical and real, the condition part of a rule in XCS combines real intervals, binary and categorical representations with their decoding as described above in Table 1. For example, the following is the first seven interval predictors in the condition part in a rule in which the first and the seventh predictors use real intervals and the others use the categorical representation: (0.0-0.9)(#)(1)(#)(#)(3)(0.0-0.3)…

An empirical investigation was performed to determine a good setting for parameters, and was found that the classification performance is sensitive to the population size N, mutation step m_0, covering step r_0 and υ. As discussed in [37], a small population size hinders the solution to be generated because of the covering and reproduction processes. Different values of population size were tested (N=5000, N=6000, N=8000, N=10000, and N=30000). And, it has been found that the population size of N=10000 provides a sufficient population size as lower values did not allow an accurate solution to evolve.

The values of r_0 and m_0 determine the initial and intermediate movements in the solution map, where r_0 is the maximum covering step size for an attribute if no rule matches the current case; and m_0 is the maximum step size that an interval can widen to while in mutation. The effect of the value of υ has been illustrated in [19].

XCS was trained using ten-fold cross validation for ten runs each over 1,000,000 iterations using roulette wheel selection with a population size of 10000. The values for all other parameters are as follows: $p_{\#}$=0.75, θ_{GA}=50, uniform crossover χ=0.8, free mutation μ=0.04, α=1, δ=0.1, ε_0=1.0, θ_{sub}=20, υ =50, r_0=0.4, m_0=0.2. Table 3 shows the classification accuracy and the average of the rulesets' size of XCS over the FBC dataset. It can be seen that XCS outperforms C4.5 and the other traditional techniques in terms of its classification accuracy which is reassuring for the capability of XCS.

Table 3. Accuracy performance for XCS.

Accuracy	Pop. size
$80.1\%_{(5.9)}$	$7974.4_{(157.4)}$

5. RULE COMPACTION

Several approaches have been attempted to develop a sufficient compaction algorithm to increases the level of rules' readability, interpretation, and organization of the underlying knowledge held in the LCS ruleset. Research in [28], [35], and [36] are some examples of these data-driven approaches which extract a minimal set of rules that covers the original dataset. However, these attempts have a common deficiency in terms of their data dependency. To select the rules that only cover the dataset will, undoubtedly, ignore a large part of the discovered domain knowledge achieved by LCS. Moreover, if the noise level of the original dataset is high, then choosing the rules that only match it will lead to noisy and inaccurate rules.

We have presented a new rule-driven compaction approach with the objective to understand LCS generated rules and their complex underlying knowledge. And, extract hidden knowledge from XCS rules by discovering interesting patterns which may highlight new domain features in addition to being efficient in classifying new future cases. Clustering has been used to produce clusters of similar rules that share most of the attributes and features. Table 4 shows the accuracy and size of the compacted ruleset obtained from the new rule-driven compaction approach. Comparisons with the data driven approaches have proven favorable in terms of size and accuracy but not in terms of quality– the reader is referred to [18] for full details and comparisons.

Table 4. Compacted performance of XCS.

RuleSet	Original	Compacted
Size	$7974.4_{(157.4)}$	$341_{(27.16)}$
Accuracy	$93.75\%_{(1..3)}$	$65.13\%_{(5.9)}$

6. DISCOVERED KNOWLEDGE

In general, the process of knowledge discovery aims to produce a novel piece of knowledge which can be easily comprehensible and useful in its problem domain. The significance of the generated knowledge can be assessed by the classification accuracy as discussed in the previous section. However, the value of the generated knowledge should also be evaluated from a domain expert point of view to emphasize if this knowledge fulfils the domain goals. That is, to evaluate the quality of the generated ruleset and to highlight the new piece of obtained knowledge (i.e., rules). This research concentrates on the medical domain, and breast cancer in particular. A medical expert, who is

a consultant pathologist in the domain of primary breast cancer, has been involved in this evaluation study to report on the relative value of the extracted and compacted knowledge from the medical point of view. Figure 2 depicts a partial extract from the evaluation forms presented to the domain expert to assess the rules generated from the previous section:

RULES	*Usefulness* 1-5	*contradiction* Y/N	*Interesting/ new knowledge* Y/N	*remarks*
. . .				
16. IF Immuno-ER-pos = TRUE and Mitotic-Figure-Score <= 1 and Histology = Ductal Carcinoma NST and Tubule-Formation-Score <= 7 THEN **Grade**=G1 (53)	1	N	N	
17. IF . . .				

Figure 2. The assessment model.

The rules have been presented as a list of if-then statements and the domain expert was kindly requested to evaluate, relying on his past experience, the medical quality of each rule individually in terms of its *usefulness, contradiction,* and whether it is *interesting* and/or represents *new knowledge*. *Usefulness* means whether the rule would be of use in predicting or classifying future cases, with a scale of 1 (if the rule is of minor useful, not discussed here) to 5 (if the rule is of the utmost significance and usefulness). *Contradiction* implies that the given rule contradicts with existing knowledge in the field or as per the domain expert's background. A rule that is marked as *Interesting* and/or *new knowledge* indicates that this rule deserves future investigation since its diagnostic knowledge seems to be either not previously highlighted or that its existence maybe suspicious. Whenever a rule is masked as interesting, the domain expert is asked to provide a brief medical explanation to verify his point view in order to enrich the output assessment.

In this investigation each rule is associated with two numbers; the number of the correctly, and the incorrectly matched cases. These two numbers certify the rule's weight and accuracy. All the rules that have not had any matched cases are dropped out and not considered as efficient throughout this evaluation study.

6.1 Analysis of C4.5 results

The randomly selected ruleset contains 85 rules in which the distribution of classes over the rules are 11%, 55% and 34% for G1, G2 and G3, respectively as illustrated in Figure 3. Based on the expert's evaluation, six rules have been found to be of high usefulness, where five rules have been considered presenting new knowledge. However, none of the rules is found to be contradicting any existing knowledge. Tables 5 and 6 present the number of interesting rules and the rules' grade of usefulness for this group of rules, respectively.

Table 7 presents the interesting rules from C4.5 according to the domain expert's evaluation. The following notes summarize the evaluation of the domain expert of the generated C4.5 rules.

In general the generated rules from C4.5 were described by the expert as a simple, easy to understand, and useful. C4.5 fails to discover some of the well-known primary cancer patterns such as the correlation between the number of involved nodes and the aggressiveness of the existing cancer.

Some rules were found too poor, maybe meaningless from the expert's point of view. For example, the rule$_i$ (*if sum >5 Then Grade=G3*), covers correctly about 95 cases (facts) and represents a naïve pattern.

It has been observed that the most useful rules in the expert's opinion match only few number of cases (facts) (i.e., between 3 and 10 cases), and that rules matched against a large number of facts seem to be not of a high value. That is, the over fitted rules seem to present a representative meaningful pattern, whereas the general rules describe useless patterns or even over general weak ones. The following rule is an example that covers more than 50 cases (facts) and has been considered as not useful at all:

IF Immuno-ER-pos = TRUE and
Mitotic-Figure-Score <= 1 and
Histology = Ductal Carcinoma NST and
Tubule-Formation-Score <= 7
THEN Grade=G1

Alternatively, the following rule is considered to be of a high value (i.e., usefulness=4) because it reveals the connection between this kind of histology and Grade 3 (G3) which can be easily used in predicting future cases, but this rule matches only three facts:

IF Immuuuno-ER-pos = FALSE and
Immuno-Done = TRUE and
LCIS-component = FALSE and
Histology = "Invasive Lobular Carcinoma" and
Report-Type=EX and
Immuni-C-erb-B2-strength=Negative
THEN Grade=G3

Furthermore, none of the rules describing Grade 1 class (G1) were found interesting as all of them are considered as useless with their usefulness category between 1 and 2 (except one rule with usefulness=3).

Table 5. C4.5: Rules' Distribution of Grade of Usefulness.

Grade of usefulness	2	3	4	5
No of rules (total number of rules=85)	8	5	6	0

Table 6. C4.5: Number of the Correctness and New/Interesting Knowledge.

	Yes
New/ interesting knowledge	5
Contradicting	0

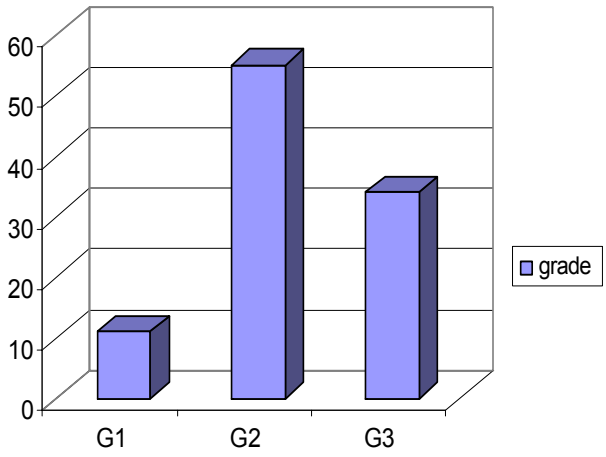

Figure 3. Class distribution for C4.5 solution.

6.2 Analysis of XCS results

A random ruleset was selected from XCS which contains a total of 2901 rules and compacted using our rule-driven compaction algorithm. Figure 4 shows the percentage of rules' distribution over the three existing classes within the XCS ruleset before compaction.

It can be seen that this distribution is balanced and each class is roughly represented equally in the ruleset generated by XCS. After applying the rule-driven compaction algorithms, the size of the compacted rulesets was 300. The compacted ruleset was then presented to the domain expert for evaluation.

For the rules compacted using the rule driven approach (300 rules), nine rules were described as interesting where ten rules were assigned to have a value 4 for usefulness. Tables 8 and 9 present the number of rules classified as useful and interesting using the rule-driven compaction, respectively. There were no contradicting rules in either results.

The expert reported that the compacted ruleset from the rule driven approach is able to find well-known patterns which described some parts of the problem domain. For example, some rules include the *Histology=Lobular Ca* and describe Grade 2 (G2) diagnosis. This type of histology refers mostly to this kind of primary cancer. The ability to find such well-known patterns assures us of the power of XCS to extract representative rules on the one hand and the rule-driven compaction approach in unveiling such hidden patterns on the other hand. Very interesting relations and interactions were identified by the expert from the generated rules especially the ones related to grade 1 (G1) class. The expert reveals that they present new knowledge regarding this type of primary cancer. Over 90% of the rules that were considered as interesting were related to Grade 1 (G1) diagnosis. In summary, these have shown that the *size of DCIS + invasive, Specimen type*, and *DCIS type* seem to be the most important attributes to determine this kind of primary breast cancer.

Table 7. Interesting rules from C4.5.

No	Condition	Grade
1	Immuno-ER-pos = FALSE Immuno-Done = TRUE LCIS-component = TRUE	G2
2	Immuno-ER-pos = FALSE Immuno-Done = TRUE LCIS-component = FALSE Histology = Invasive Lobular Carcinoma Report-Type = EX Immuno-C-erb-B2-strength = Negative	G3
3	Immuno-ER-pos = FALSE Immuno-Done = FALSE DCIS-Necrosis = TRUE THEN Grade	G3
4	Immuno-ER-pos = FALSE Immuno-Done = FALSE DCIS-Necrosis = FALSE Histology = Ductal Carcinoma NST age <= 39	G3
5	Immuno-ER-pos = FALSE Immuno-Done = FALSE DCIS-Necrosis = FALSE Histology = Ductal Carcinoma NST age > 39	G2

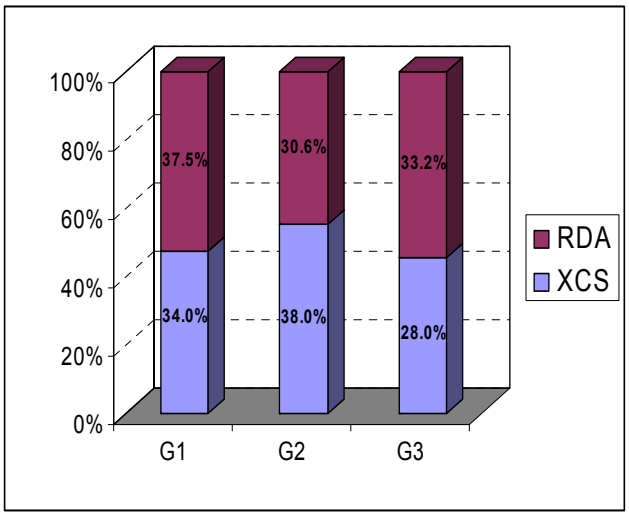

Figure 4. Percentage of Rules' Distribution over the G1, G2, and G3 Classes in XCS and Rule Driven Approach to compaction in a randomly selected ruleset.

Table 8. Rule-Driven Compaction: Rules' Distribution of Grade of Usefulness.

Grade of usefulness	2	3	4
No of Rules (total number of rules=300)	50	5	10

Table 9. **Rule-Driven Compaction: Number of the Correctness and New/Interesting Knowledge.**

	Yes
New/ interesting knowledge	9
Contradicting	0

Some of the generated patterns were found inappropriate for use, as their left hand side consists of a high number of attributes. From the domain expert's point of view, it is very difficult to utilise rules having a large number of participating attributes in the real domain application. Therefore, this criterion may prevent many rules from being considered as interesting or useful. Moreover, the encoding and clustering technique led to some patterns with the following action: "*G1 or G2 or G3*" which represents a conflict pattern of symptoms and attributes. Such patterns are also not appropriate to be used in real applications.

Overall, the rules were judged to be more informative and qualitatively useful than the ruleset obtained from C4.5. They were found to contain more rich and descriptive patterns that cover the problem space more widely.

The expert revealed that because XCS generates larger rulesets, more time is needed to search for such unique, surprising, and newly hypothesised rules. This may be considered a barrier to arrive at the maximum benefit from the generated rulesets.

Although the classification accuracy of Grade 3 (G3) diagnoses is 91%, the expert does not find the rules covering this category interesting nor contributing to new knowledge. However, this may be attributed to the expert's experience being based on the cases collected from different patients and from his ongoing research. Grade 3 (G3) is the most aggressive primary cancer, on which more research is being carried out, more knowledge is evolving, and attention is given to provide a better understanding of this type of cancer, its causes and diagnosis. Therefore, although the generated rules cover over 90% of the cases (facts), it is suggested that the expert's expectation is higher than the knowledge hidden within the given cases (facts). Therefore, most of the generated rules that cover Grade 3 (G3) diagnosis were considered not interesting. Space restrictions prohibit showing the nine most interesting rules learned by XCS (see [19], available online).

7. CONCLUSION

This paper has reported on the knowledge discovery process applied to medical databases and, in particular, the area of primary breast cancer. Results indicate that using a learning classifier system, such as XCS, followed by the appropriate compaction process, can lead to knowledge discovery and rule induction from such datasets. The process started with preparing the data in terms of rule encoding and then balancing the representation of the involved data classes.

In addition, the dataset was applied to different learning techniques. XCS outperformed C4.5 and other selected learning techniques. The rulesets generated from XCS was compacted using rule driven approach and was qualitatively evaluated by a domain expert using the pre-defined criterion. The same criterion was used to evaluate the ruleset induced from C4.5 to allow the comparison to be carried out with rulesets from XCS.

In summary, rules evolved by XCS were qualitatively comparable to those induced from C4.5. Although C4.5 produces fewer rules, XCS was found to breed more useful descriptive ones. In the case of the C4.5 ruleset, the expert recognized some missing patterns and pointed out some poor quality rules. In XCS, rules are richer and representative but more complicated as well.

The involvement of a domain expert specialist who evaluated the rules relying on his previous knowledge guided by the assessment model enabled critical evaluation of the final rulesets to assess the quality of the rules generated from different perspectives. This involvement can be extended with more complicated assessment models to be utilized in refining the compaction technique or any related issue.

8. REFERENCES

[1] Bacardit, J., & Butz, M. (2004). Data Mining in Learning Classifier Systems: Comparing XCS with GAssis. *Advances in Learning Classifier Systems (7th International Workshop, IWLCS 2004)*, Seattle, USA, LNAI, Springer.

[2] Batista, G., Prati, R., & Monard, M. (2004). A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations, 6(1), pp 20-29.

[3] Bernado, E., Llorà , X., & Garrell, J. (2002). XCS and GALE: a Comparative Study of Two Learning Classifier Systems on Data Mining. *Advances in Learning Classifier Systems, 4th International Workshop, volume 2321 of Lecture Notes in Artificial Intelligence,* Springer, pp 115-132.

[4] Blake, C., & Merz, C. (1998). UCI Repository of Machine Learning Databases [online]. Irvine, CA: University of California, Department of Information and Computer Science. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html [Accessed 2/2004].

[5] Bonelli, P. & Parodi (1991) An Efficient Classifier System and its Experimental Comparison with Two Representative Learning Methods on Three Medical Domains. In *Proceedings of the 4th International Conference on Genetic Algorithms*, pp. 288-295, Morgan Kauffman.

[6] Butz, M. & Wilson, S.W. (2001) An Algorithmic Description of XCS. In *Advances in Learning Classifier Systems: Proceedings of the Third International Conference – IWLCS2000*. Springer, pp. 253-272.

[7] Famili, F., Shen, W., Weber, R., & Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. *Intelligent. Data Analysis,* 1(1-4), pp 3-23.

[8] Holland, J.H. (1986). Escaping brittleness: the possibilities of general-purpose learning algorithms applied to parallel rule-based systems. In Michalski, Carbonell, & Mitchell (eds) *Machine learning, an artificial intelligence approach*. Morgan Kaufmann.

[9] Holmes, J. (1997). Discovering risk of disease with a learning classifier system, In: T. Baeck (ed). *Proceedings*

of the Seventh International Conference on Genetic Algorithms (ICGA97). Morgan Kaufmann, San Francisco, CA (1997).

[10] Holmes, J. (2000). Learning Classifier Systems Applied to Knowledge Discovery in Clinical Research Databases. In: P. Lanzi, W. Stolzmann, S. W. Wilson (eds). *Learning Classifier Systems: From Foundations to Applications.* Springer-Verlag Heidelberg, pp 243-261

[11] Holmes J., Bilker W., (2002) The Effect of Missing Data on Learning Classifier System Learning Rate and Classification Performance. IWLCS 2002: 46-60.

[12] Holmes, J., Sager, J., & Bilker, W. (2004). A Comparison of Three Methods for Covering Missing Data in XCS. Seventh International Workshop on Learning Classifier Systems (IWLCS-2004).

[13] Japkowicz, N. (2003). Class Imbalances: Are we focusing on the Right Issue? Notes from the ICML Workshop on Learning from Imbalanced Data Sets II.

[14] Japkowicz, N., and Stephen, S.(2002) The class imbalance problem: a systematic study. Intelligent Data Analysis, 6(5):429-450.

[15] Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI '2000).

[16] Jensen, F. (1996). An Introduction to Bayesian Networks. Springer.

[17] Jo, T., & Japkowicz, N. (2004). Class Imbalances versus Small Disjoints, *SIGKDD Explorations,6(1).*

[18] Kharbat, F. (2006) Learning Classifier Systems for Knowledge Discovery in Breast Cancer. PhD Dissertation. University of the West of England, U.K.

[19] Kharbat, F., Bull, L. & Odeh, M. (2005) Revisiting Genetic Selection in the XCS Learning Classifier System. In *Proceedings of the IEEE Congress on Evolutionary Computation*. IEEE, pp2061-2068.

[20] Kohavi R. & Provost F. (1998)., Glossary of Terms. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. 30: 2/3.

[21] Liu, W., Shen, P., Qu, Y., & Xia, D. (2001). Fast algorithm of support vector machines in lung cancer diagnosis. International Workshop on Medical Imaging and Augmented Reality, 2001, pp 188-192.

[22] Moore, A., & Hoang, A. (2002). A performance assessment of Bayesian networks as a predictor of breast cancer survival. Second international workshop on Intelligent systems design and application, pp 3 – 8.

[23] Ogunyemi, O., Chlebowski, R., Matloff, E., Schnabel, F., Orr, R., & Col, N. (2004). Creating Bayesian Network Models for Breast Cancer Risk Prediction. Cancer Risk Prediction Models: A Workshop on Development, Evaluation, and Application, pp 20-21.

[24] Quinlan, J. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

[25] Sierra, B., & Larranaga, P. (1998). Predicting the survival in malignant skin melanoma using Bayesian networks. An empirical comparison between different approaches, Artificial Intelligence in Medicine, 14(1-2), pp 215-230.

[26] Sorace, J., & Zhan, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 4(24).

[27] Tan, K., Yu, Q., Heng, C., & Lee, T. (2003). Evolutionary computing for knowledge discovery in medical diagnosis, *Artificial Intelligence in Medicine*, 27(2), pp 129-154.

[28] Wilson, S. (2001). Compact Rulesets from XCSI. *Fourth International Workshop on Learning Classifier Systems* (IWLCS-2001). pp. 197-210, San Francisco, CA

[29] Wilson, S., (1995) Classifier Fitness Based on Accuracy. *Evolutionary Computing* **3**: 149-175.

[30] Witten, I., & Frank, E., (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco

[31] Wyatt, D., Bull, L., & Parmee, I. (2004). Building Compact Rulesets for Describing Continuous-Valued Problem Spaces Using a Learning Classifier System. In: I. Parmee (ed). Adaptive Computing in Design and Manufacture VI. Springer, pp 235-248.

[32] Prati, R.., Batista, G., and Monard, M. (2004) Class Imbalances versus Class Overlapping: an Analysis of a Learning System Behavior. In MICAI, pp. 312-321.

[33] Barandela, R., Valdovinos, R., Scnchez, J., & Ferri, F. (2004). The imbalanced training sample problem: Under or over sampling?. *Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition* (SSPR/SPR'04). Lecture Notes in Computer Science 3138, Springer-Verlag, pp 806-814.

[34] Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning In*: B. Schoelkopf, C. Burges, and A. Smola (eds). MIT Press, pp 185 – 208.

[35] Fu, C., & Davis, L. (2002). A Modified Classifier System Compaction Algorithm. *Genetic and evolutionary computation conference (GECCO-2002)*, pp 920-925.

[36] Dixon, P., Corne, D., & Oates, M. (2003). A Ruleset Reduction Algorithm for the XCS Learning Classifier System. In: P. Lanzi, W. Stolzmann, and S. Wilson (eds.) *Proceedings of the 3rd International Workshop on Learning Classifier Systems*. Springer LNCS, pp 20-29.

[37] Butz, M., Goldberg, D., & Tharakunnel, K. (2003). Analysis and improvement of fitness exploration in XCS: bounding models, Tournament selection and bilateral accuracy. *Evolutionary Computation*, 11(3), pp 239-277.