ARC-UI: A Visualization Tool for Associative Classifiers

David Chodos, Osmar Zaïane
Department of Computing Science, University of Alberta, Canada
chodos@cs.ualberta.ca, zaiane@cs.ualberta.ca

Abstract

The classification of an unknown item based on a training data set is a key data mining task. An important part of this process that is often overlooked is the user's comprehension of the classifier and the results it produces. Associative classifiers begin to address this issue by using sets of simple rules to classify items. However, the size of these rule sets can be an obstacle to understandability. In this work, we present an interactive visualization system that allows the user to visualize various aspects of the classifier's decision process. This system shows the rules that are relevant to the classification of an item, the ways in which the item's characteristics relate to these rules, and connections between the item and the classifier's training data set. The system also contains a speculation component, which allows the user to modify rules within the classifier, and see the impact of these changes. Thus, this component allows the user to contribute domain expertise to the classification process, consequently improving the accuracy of the classifier.

Keywords— **visualization, associative classifiers, classification result analysis**

1 Introduction

The classification of items based on previously classified training data is an important area within data mining, and has many real-world applications. However, one drawback to many classification techniques, such as those based on neural networks or support vector machines (SVM), is that it is difficult for the user to understand the reasoning behind a classification result, or interpret the learned classification model. This is particularly important in a context where an expert user could make use of domain knowledge to either confirm or correct a dubious classification result.

Rule-based classifiers address this shortcoming by using a collection of simple rules to perform classification. Each rule is made up of one or more attribute/value pairs and a class, and is thus quite easy to understand. Most rule-based classifiers perform a heuristic search to discover classification rules, often missing important ones. Associative classifiers [12, 11], on the other hand, use association rule mining [1] to perform an exhaustive search to find classification rules. However, the set of rules generated by an associative classifier may contain hundreds of thousands of rules, and thus it is difficult for the user to ascertain which rules are relevant to the classification of an item, and to what extent the relevant rules influence a classification decision.

This paper presents ARC-UI, a tool that allows the user to understand the reasoning behind an associative classification result via a graphical, interactive interface. Although other rule visualizers exist [6, 9], ARC-UI is unique in that the user is able to modify the rules that are used and immediately see the results of this modification, thus allowing the user to improve the accuracy of the classifier through the application of domain expertise. This capability has the added benefit of increasing the user's confidence in the classifier.

2 Related Work

Research related to visualizing associative classification results can be divided into three areas: visualizing classification results, visualizing association rule sets, and visual classification. Each of these areas contributes important ideas and techniques to the aims of this research.

2.1 Visualizing Classification Results

One of the main goals of this research is to allow users to understand the result of classifying an item. Poulin, Szafron and others have investigated this issue of classification result analysis [13, 15]. Although their work focuses on additive, rather than rule-based, classifiers, the concerns that are identified are equally applicable to this research. An additive classifier uses a sum of terms to classify an item, where each term represents the likelihood of an attribute belonging to a given class, and the sum indicates the likelihood of the item belonging to that class. Each term in the equation has a weight associated with it, indicating its relevance to the classification decision [13]. Classifiers which fall into this category include Bayesian classifiers, SVM-based classifiers, and linear regression classifiers [8]. Their work is quite relevant, in that they present a system which allows users to analyze classification results. The

1550-6037/08 $25.00 © 2008 IEEE
DOI 10.1109/IV.2008.35

authors propose five desired areas for analysis:

- Classification: Show the classification decision made by the classifier, and show the alternatives.
- Decision evidence: Show the evidence that was used by the classifier to arrive at its result.
- Decision speculation: Show the effect of changing the item that is being classified.
- Ranks of evidence: Show the relative importance of the evidence used by the classifier.
- Source of evidence: Show the data that was used by the classifier to create the classification model.

The authors also present a system, ExplainD, which fulfills these requirements for additive classifiers. For each of these classifiers, the authors show how ExplainD implements the five analysis areas described above.

This paper was followed by work by Szafron *et al.* which applied the ideas of ExplainD to the problem of proteome analysis [15]. This system was able to provide good prediction results in a variety of proteins, and make "every prediction transparent to its users" through its explanatory features. In their decision speculation component, they only change the attribute value of the object being classified, and do not edit any part of the learned model. Although this system still relies on additive classifiers, its success provides real-world evidence that classification result analysis is a useful capability, and well worth pursuing.

2.2 Visualizing Association Rule Sets

A major drawback of associative classifiers is that they use very large rule sets to perform classification. Thus, understanding of the classification result is hampered by the need to make sense of sets of hundreds or thousands of rules. This issue is particularly relevant to this research because of the decision evidence component which, in an associative classification context, relies on the visualization of the rules which are used in classifying an item. Several researchers have addressed this issue in the context of association rule mining.

One method of dealing with large associative rule sets was developed recently by Tuzhilin and Adomavicius within the context of microarray data. They developed post-processing techniques, such as grouping and filtering, which made the analysis of very large numbers of association rules a more manageable task [16]. As this work is focused more on processing methods than visualization, it is not directly applicable to this research. However, the grouping techniques may be useful in implementing visualization techniques, especially when considered along with the work of Couturier, which is described below.

Another step in this direction was taken by Rahal *et al.* Their solution, however, focuses on querying this rule set in order to find "the subset of associations that are of interest [in an] interactive mode" [14]. The idea of allowing the user to interactively analyze association rules is similar in spirit to the visual classification work by Ankerst, described in the following section. However, the need to iteratively refine the rule set conflicts with the need, in the context of this research, to quickly present the user with a summary of a large rule set.

Fukuda and Morimoto developed an interesting visualization technique for optimized two-dimensional rules [7]. Their work focuses on rules where the domains of the attributes form a planar region. For these rules, the authors represent the region as a pixel map, where each pixel is assigned a colour and brightness level which convey information about that point in the plane. This work is intriguing in its use of visual elements such as position, brightness and colour to represent association rules. However, it is limited to rules of a particular form, and thus is not practical for general-purpose analysis.

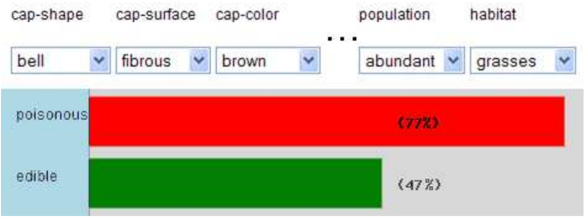
Couturier *et al.* have also investigated visualizing large sets of association rules [5]. Rather than trying to represent each rule individually, the authors propose clustering the rules, and then using a fish-eye view (FEV) technique to view the details of a particular cluster while viewing coarse-grained representations of other clusters.

Leung *et al.*, in developing a visualization system for frequent itemsets, analyzed a couple of creative methods for visualizing association rules [10]. One of these, AViz, uses a planar representation method similar to that developed by Fukuda and Morimoto. The other, by Yang, uses a Bezier curve to represent a rule, thus allowing multiple rules, each represented by a Bezier curve, to be shown on the same graph. These techniques would not be effective for large rule sets, but could be helpful for visualizing small subsets of association rules.

2.3 Visual Classification

Ankerst *et al.* investigated the concept of having the user assist in the process of creating the classifier through visual inspection of the data, and evaluation of the classifier as it is created [2]. This research is based on the idea that the user can contribute domain knowledge and pattern recognition abilities to the classifier creation process, both of which are valuable contributions.

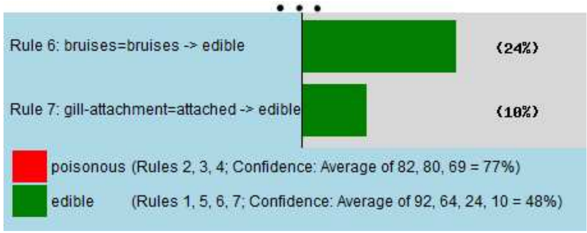
While the work focuses on decision tree-based classifiers, the visualization techniques presented offer innovative ways of effectively presenting large quantities of data. Furthermore, the research emphasizes the importance of collaboration between the user and the classification system in creating an effective classification model, which



cap-shape: bell
cap-surface: fibrous
cap-color: brown
. . .
population: abundant
habitat: grasses

poisonous (77%)
edible (47%)

Figure 1: Classification component



. . .

Rule 6: bruises=bruises -> edible (24%)
Rule 7: gill-attachment=attached -> edible (10%)

🟥 poisonous (Rules 2, 3, 4; Confidence: Average of 82, 80, 69 = 77%)
🟩 edible (Rules 1, 5, 6, 7; Confidence: Average of 92, 64, 24, 10 = 48%)

Figure 2: Decision evidence component (truncated)

echoes the goals of our research.

3 System Features

The initial development of the system was guided by the five analysis components described by Poulin *et al.* in their work on the ExplainD system [13]. Through testing and user feedback, other requirements were identified, which led to further development. These components had already been implemented in a linear classification context, but our focus on associative classification meant that some of the components had to be significantly revised.

The screenshots that follow were taken from system's use in the context of classifying mushrooms. The well-known "mushroom" data set, downloaded from the UCI data repository, contains over 8,000 mushrooms that have been classified as either poisionous or edible [4]. Each item in the data set contains twenty-two characteristics, such as gills-attached, colour and odor, that help determine the item's classification. The Weka data analysis tool [17] was used to generate classification rules (1,275 in total), which were then imported into the system. Thus, the screenshots show the system analyzing the classification of an unknown mushroom using these rules.

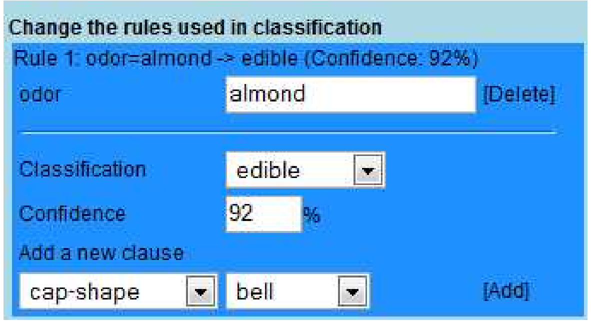
The classification component shows the result of classifying the item, as well as all other possible classifications, as shown in Figure 1. This allows the user to compare the result with other possibilities, and thus assess the likelihood of an alternative result. The classification possibilities are listed in decreasing order of likelihood, to facilitate comparison between the various possibilities.

The decision evidence component shows the rules that were used to classify an item. This gives the user an initial understanding of the reasoning used by the classifier. If the relevant rule set is small enough, these rules are shown in a bar graph, as in Figure 2. However, if the rule set is too large for this to be feasible (i.e., more than a few dozen rules), the bar graph is compressed in order to present the rule set characteristics in a meaningful, visual manner. Using this "compressed" format, hundreds of rules may be viewed in a summarized form. In either case, the bar graph is colour-coded according to the class labels, to facilitate comparison among the rules shown. As well, the component presents a summary of the rules influencing each classification possibility. This summary, which is generated by sorting the relevant rules according to their classification labels, includes the confidence value for each rules and the overall confidence for each class. The overall confidence is calculated using either the best rule [12] or average rule method [3], as specified by the user.

The decision speculation component allows the user to modify the item being classified, the method used to calculate the confidence for each class, and the rules used in classification. The user is shown a list of all relevant rules, and can click on any rule to view a simple editing menu for that rule, shown in Figure 3. This allows the user to deal with a large set of potentially relevant rules, while providing fine-grained editing capabilities where they are needed. After performing the desired modifications, the user is immediately shown the results of this modification. This allows the user to experiment with the classification engine, thus offering insight into the process behind item classification. In selecting the confidence calculation method, the user may choose between the best rule [12] and average rule methods [3]. When editing the rules used in classification, the user can:

- edit the classification or confidence for a rule
- add, modify or delete clauses within a rule
- remove a rule entirely (causing it to be ignored)
- create a new rule

Thus, the user can draw on expert knowledge to edit the computationally-generated rule set. Moreover, the user is shown immediately whether this modification improved the accuracy of the classifier. It should be noted that the speculative changes made by the user are not immediately made permanent. However, the user has the option of making the speculative changes permanent in the classification model, once the results of these changes have been presented, and accepted by the user. Thus, the tool offers the ability to interactively analyze and improve the classifier.



Change the rules used in classification

Rule 1: odor=almond -> edible (Confidence: 92%)

odor: almond [Delete]

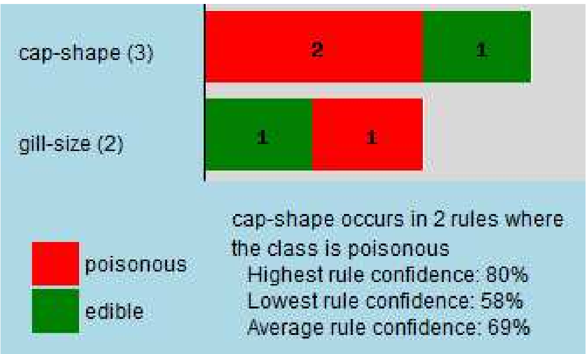
Classification: edible

Confidence: 92 %

Add a new clause

cap-shape bell [Add]

Figure 3: Decision speculation component



cap-shape occurs in 2 rules where the class is poisonous

Highest rule confidence: 80%

Lowest rule confidence: 58%

Average rule confidence: 69%

Figure 4: Ranks of evidence component

The ranks of evidence component shows the relationships between characteristics, association rules and classifications. This provides the user with further information about the way the classifier works, independent of any particular item or rule set. The system uses a colour-coded bar chart-based visualization scheme, as shown in Figure 4. The length of each bar indicates the total number of times a characteristic appears in the rule set. The colour-coded segments show the number of rules containing a given characteristic that result in a particular classification. By moving the mouse over each segment, the use is shown a more detailed summary of the rules that contain a given characteristic and result in the selected classification. This approach is both visually appealing and scalable, which is quite beneficial when dealing with very large rule sets. In Figure 4, we see that the "cap-shape" characteristic appears in three rules, two of which have the class "poisonous", and one with the class "edible" (represented by green and red segments, respectively). By placing the mouse over the "poisonous" segment of the bar for the "cap-shape" characteristic, we are shown more information about the rules containing the "cap-shape" characteristic where the class is "poisonous".

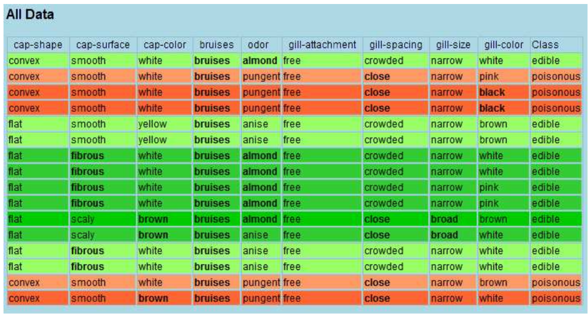
Finally, the source of evidence component allows the user to make connections between the item being classified and the entries in the data set that were used to generate the associative classification rules. This may be useful when investigating a dubious classification result - the user can check the data used in training the classifier to see if there were any anomalies in that original data. Specifically, the component shows the entries in the training set in a colour-coded list using shades of red and green, as shown in Figure 5. A green entry indicates that the entry has the same class as the item being analyzed, while a red entry indicates that they have different classes. The intensity of the colour indicates the proximity of the entry to the current item, in terms of matching attribute-value pairs. Finally, the user is able to specify a variety of further analysis options to restrict the list of entries to those matching certain classification or characteristic criteria. In particular, when filtering by attribute, the user is shown a chart of the distribution of that attribute among the possible classifications, divided by the possible attribute values. Figure 6 shows the class breakdown for the possible values of the "cap-shape" attribute. For example, in 81% of the 535 items containing the "cap-shape=convex" attribute-value pair, the class was "edible". The table also shows that there were no items which contained "cap-shape=conical", and thus this attribute-value pair had no impact on the classification model.

4 Evaluation

As described in the methodology section, the system was evaluated by testing it with several real-world data sets, and assessing the system's ability to analyze classification results, deal with large sets of association rules, and provide the user with interactive rule set editing capabilities.

Furthermore, one of the goals of the system is to allow the user to use their expertise to interactively modify the rule set used by the classifier, and thus improve the overall accuracy of the classifier. Showing that this is indeed possible is another form of validation for the system as a whole.

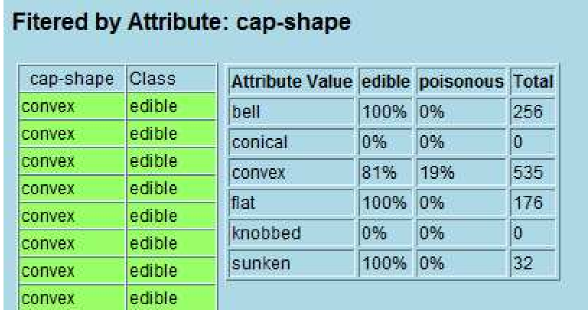
To perform the first part of the validation, several data sets were obtained from the UCI machine learning repository mushroom, car evaluation, and nursery [4]. These data sets were then fed into Weka [17], a data analysis program, and the resulting rules were then pruned for relevance. For a data set with n classes, rules with a confidence value less than $\frac{1}{n}$ were pruned, since they indicated less certainty than would be provided by random selection. Finally, the relevant rules, along with the original data sets, were loaded into our system. The data sets had varying numbers of items and attributes, and the rule sets were similarly varied, thus providing a good test of the system's ca-



All Data

cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	Class
convex	smooth	white	**bruises**	**almond**	free	crowded	narrow	white	edible
convex	smooth	white	**bruises**	pungent	free	**close**	narrow	pink	poisonous
convex	smooth	white	**bruises**	pungent	free	**close**	narrow	**black**	poisonous
convex	smooth	white	**bruises**	pungent	free	**close**	narrow	**black**	poisonous
flat	smooth	yellow	**bruises**	anise	free	crowded	narrow	brown	edible
flat	smooth	yellow	**bruises**	anise	free	crowded	narrow	brown	edible
flat	**fibrous**	white	**bruises**	**almond**	free	crowded	narrow	white	edible
flat	**fibrous**	white	**bruises**	**almond**	free	crowded	narrow	white	edible
flat	**fibrous**	white	**bruises**	**almond**	free	crowded	narrow	pink	edible
flat	**fibrous**	white	**bruises**	**almond**	free	crowded	narrow	pink	edible
flat	scaly	**brown**	**bruises**	**almond**	free	**close**	**broad**	brown	edible
flat	scaly	**brown**	**bruises**	anise	free	**close**	**broad**	white	edible
flat	**fibrous**	white	**bruises**	anise	free	crowded	narrow	white	edible
flat	**fibrous**	white	**bruises**	anise	free	crowded	narrow	white	edible
convex	smooth	white	**bruises**	pungent	free	**close**	narrow	brown	poisonous
convex	smooth	**brown**	**bruises**	pungent	free	**close**	narrow	white	poisonous

Figure 5: Source of evidence component



Fitered by Attribute: cap-shape

cap-shape	Class
convex	edible
convex	edible
convex	edible
convex	edible
convex	edible
convex	edible
convex	edible
convex	edible

Attribute Value	edible	poisonous	Total
bell	100%	0%	256
conical	0%	0%	0
convex	81%	19%	535
flat	100%	0%	176
knobbed	0%	0%	0
sunken	100%	0%	32

Figure 6: Source of evidence chart

pabilities (see Table 1 for details).

In order to test the utility of the interactive classifier modification component, each data set was analyzed and iteratively improved using the decision speculation component. In a scheme similar to the k-fold cross-validation method of classifier validation, 90% of the items in each data set were used to train a classifier, while the remaining 10% of the items were set aside for use with the ARC-UI system. For each of these items, the ARC-UI system was used to classify the item, and the result was compared to the correct class, as recorded in the original data set. The speculation component of the system was then used to modify the classifier and test the modified classifier, with the aim of fixing the errors made originally. Finally, after modifications were made so that the incorrectly classified items were classified correctly, all of the items were reclassified using the modified classifier, to assess the overall impact of the modifications and ensure that the changes had not adversely affected the accuracy of the classifier. The results of classification using the initial and modified classifiers are presented in Table 2.

As shown in this table, the system was used to improve the accuracy of each classifier, using the speculation and classification tools. Furthermore, it is worth noting that, in each case, the increased accuracy was caused by a very small change in the rule set. This indicates that, by identifying the appropriate rules to modify, the user can improve the accuracy of the classifier with a minimal amount of time and effort.

5 Future Work

There are several promising areas for further work on this project. Three of these are empirical validation, rule set visualization, and the improvement of the source of evidence component.

The validation for the system has, thus far, consisted of ensuring that the system met its functional requirements, and offered users the ability to interactively improve the classifier they are working with. However, a key non-functional requirement of the system is that users from a broad range of backgrounds - that is, beyond computer science and data analysis - should find the system intuitive and easy to use. To find out whether this is, in fact, the case, it would be beneficial to conduct a study to investigate the system's usability. Participants could be drawn from a range of subject areas, and thus the experiment could measure the system's applicability to specific contexts, as well as its general usability.

Effective rule set visualization becomes challenging when the size of the relevant rule set exceeds a few dozen rules. Currently, the system presents a compressed bar graph visualization if the rule set is larger than a certain threshold. However, it may be more effective to use a clustering approach, similar to that proposed by Couturier *et al.* [5], or a visualization scheme inspired by those presented by Ankerst *et al* [2]. Thus, several visualization options could be implemented and then tested with users drawn from the participant pool described previously.

Currently, the source of evidence helps the user identify anomalous entries in the original data set. It would be even more beneficial if the user could, after identifying such entries, modify or even remove them entirely. This, along with an integrated classification rule generator, would allow the user to interactively fix errors in the training data.

The analysis techniques currently implemented are based on those developed for the linear classifier-oriented system by Poulin. However, there may be other analysis techniques which would be particularly useful for an associative classification context or other rule-based models.

Conclusions

In this paper, we present ARC-UI, an interactive system for analyzing the results of an associative classifier. Associative classifiers have the advantage of easily understandable classification rules, but often use very large rule sets. The system offers a variety of tools to help the user understand the classification engine's result, including visualizations of relevant rule sets; the relationships between characteristics, rules and classes; and connections between training data and the item being classified.

Name	Items	Attrs	All rules	Pruned rules
Car	1,728	6	1,057	28
Mushroom	8,124	22	1,275	71
Nursery	12,960	8	2,418	10

Table 1: Data set characteristics

	Car	Mushroom	Nursery
Number of items	173	813	130
Rules modified	2	1	1
Accuracy (before)	80.9%	93.5%	70.4%
Accuracy (after)	82.1%	94.8%	73.8%

Table 2: Initial vs. modified classifier accuracy

Perhaps most important, however, is the speculation component, which allows the user to modify the rules used to classify an item, and then immediately see the results of classifying the item using the modified rule set. Thus, the user can contribute domain knowledge to the classification process, improving the classifier's accuracy and increasing the user's confidence in the reasoning behind the classifier's decision-making process.

The system was validated by assessing its effectiveness with several data sets drawn from a variety of contexts, and also by testing the utility of the speculation component described previously. Thus, it was shown that the system could be used to understand the results of classifying items in various real-world contexts, and that the speculation tool could be used to improve the classifier's accuracy.

Acknowledgements

This research has been funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.

[2] M. Ankerst, M. Ester, and H. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *KDD 2000*, pages 179–188, 2000.

[3] M. Antonie and O. Zaïane. Text document categorization by term association. *IEEE Data Mining (ICDM)*, pages 19–26, 2002.

[4] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[5] O. Couturier, T. Hamrouni, S. Ben Yahia, and E. Mephu Nguifo. A scalable association rule visualization towards displaying large amounts of knowledge. In *IV '07: Conf. on Information Visualization*, pages 657–663, 2007.

[6] Usama Fayyad, Georges G. Grinstein, and Andreas Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2002.

[7] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining with optimized two-dimensional association rules. *ACM Trans. Database Syst.*, 26(2), 2001.

[8] J. Han and M. Kamber. *Data Mining, Concepts and Techniques, 2nd edition*. Morgan Kaufmann, 2006.

[9] TuBao Ho, TrongDung Nguyen, and DucDung Nguyen. Visualization support for a user-centered kdd process. In *SIGKDD '02*, pages 519–524, 2002.

[10] C.K. Leung and C. Carmichael. Frequent itemset visualization. Technical report, U. of Manitoba, 2007.

[11] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classifcation based on multiple class-association rules. *ICDM*, pages 369–376, 2001.

[12] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. *Proc. of SIGKDD*, pages 80–86, 1998.

[13] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D.S. Wishart, A. Fyshe, B. Pearcy, C. MacDonnell, and J. Anvik. Visual explanation of evidence in additive classifiers. In *Proc. Conf. on Innovative Applications of Artificial Intelligence*, pages 1–8, 2006.

[14] I. Rahal, D. Ren, A. Perera, H. Najadat, W. Perrizo, R. Rahhal, and W. Valdivia. Incremental interactive mining of constrained association rules from biological annotation data with nominal features. In *SAC '05: ACM Symposium on Applied computing*, 2005.

[15] D. Szafron, P. Lu, R. Greiner, D. Wishart, Z. Lu, B. Poulin, R. Eisner, J. Anvik, and C. MacDonnell. Proteome analyst - transparent high-throughput protein annotation: Function, localization and custom predictors. Technical report, U. of Alberta, 2003.

[16] A. Tuzhilin and G. Adomavicius. Handling very large numbers of association rules in the analysis of microarray data. In *Proc. ACM SIGKDD*, 2002.

[17] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman, 2nd edition edition, 2005.