Informative Sampling for Large Unbalanced Data Sets

Zhenyu Lu[1], Anand I. Rughani[2], Bruce I. Tranmer[2], Josh Bongard[1]
[1]Department of Computer Science
[2]Division of Neurosurgery
University of Vermont
Burlington, VT 05401
zlu@uvm.edu

ABSTRACT

Selective sampling is a form of active learning which can reduce the cost of training by only drawing informative data points into the training set. This selected training set is expected to contain more information for modeling compared to random sampling, thus making modeling faster and more accurate. We introduce a novel approach to selective sampling, which is derived from the Estimation-Exploration Algorithm (EEA). The EEA is a coevolutionary algorithm that uses model disagreement to determine the significance of a training datum, and evolves a set of models only on the selected data. The algorithm in this paper trains a population of Artificial Neural Networks (ANN) on the training set, and uses their disagreement to seek new data for the training set. A medical data set called the National Trauma Data Bank (NTDB) is used to test the algorithm. Experiments show that the algorithm outperforms the equivalent algorithm using randomly-selected data and sampling evenly from each class. Finally, the selected training data reveals which features most affect outcome, allowing for both improved modeling and understanding of the processes that gave rise to the data.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms

Keywords

sampling, active learning, coevolution, EEA

1. INTRODUCTION

Each year 1.4 million Americans sustain traumatic brain injury (TBI) [15]. While this leads to 50,000 deaths and 235,000 hospitalizations, the outcomes are far from uniform. Accurately predicting outcomes from head injury is a significant clinical challenge, and has implications that extend from treatment decisions to family counseling [20]. Factors reported to affect the prognosis in severe traumatic brain injury are myriad and include, for example, Glasgow Coma Scale (GCS) score, age, pupillary response and size, hypoxia, hyperthermia, and high intracranial pressure [12].The complex interactions of multiple variables contribute to the difficulty in providing accurate prognosis early in TBI. Recently, developments in data mining have made it possible to build models from such large data sets with non-uniform outcome distributions.

Classification is a technique for building prediction models from known data with both input and label information [30][24]. In many real world applications, determining the label for a given data point is sometimes expensive because it requires human experts' time. For a very large data set, further difficulties arise because even when labels are easy to obtain, it is sometimes infeasible to use all data points for modeling. Choosing a small subset of data points for labeling and then modeling therefore is of great interest.

Selective sampling is an active learning technique that selects only the informative data points for labeling, and then uses them for creating classification models. This technique renders the modeling task faster and less expensive while retaining accuracy. Many selective sampling methods have been reported. Uncertainty sampling [16] chooses the data points for which a classifier is least confident. Query by committee [28] selects the data points that cause maximal disagreement among a set of candidate models. [22] trains different classifiers on different views (disjoint subsets of features that can be used for learning), and uses disagreement among classifiers to select data. [13] uses variation in label assignments (of unlabeled data) between the classifier trained on the known training set and the classifiers training on the set with a single unlabeled object added with all possible labels.

The target data set is a medical data set called the National Trauma Data Bank (NTDB). NTDB is established by the American College of Surgeons (ACS)[1] as a public service for the trauma care community. It is a very large (over 1 million data points) and highly unbalanced data set in terms of mortality. The characterization of this data set has made some previous selective sampling methods difficult to apply. [22] requires multiple views for training purposes which re-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
Copyright 2008 ACM 978-1-60558-130-9/08/07 ...$5.00.

[1] Information and resources related to NTDB can be found at http://www.facs.org/trauma/ntdb.html

quires information that is not available from NTDB. [13] needs to scan the whole unlabeled data set once for each chosen data point. For such a large data set as NTDB, this is prohibitively time consuming. Uncertainty sampling needs to scan the data set each time a new data point is added to find the least confident data. It also requires a legitimate confidence measure for the classifier, which is not always easy to determine. Query by committee suggests an interesting way to choose data points, but the method to develop diversified yet accurate models and disagreement-causing tests is left open. The estimation-exploration algorithm[1] is a general algorithm that builds on query by committee: it actively requests the data points it wants and uses them for modeling. What distinguishes it from query by committee is that an evolutionary algorithm is used to optimize models, and another evolutionary algorithm is used to evolve tests to induce model disagreement. It has proven successful in evolutionary robotics [4], grammatical inference [2] and nonlinear systems modeling [3].

In this paper, the EEA is applied to NTDB as a selective sampling method, which is called informative sampling. The algorithm works by evolving the set of candidate models and candidate tests in an iterative manner. In each round of the algorithm, a data point is chosen based on the disagreement it causes within the current set of candidate models, and is added to the training set. Then the candidate models are trained on the updated training set. The coevolution of these two components often accelerates modeling. Importantly, the algorithm models large data sets with only one scan through the data set.

Many classification algorithms have been suggested such as decision trees [24], artificial neural networks [30] and support vector machines [6]. While the majority of them make the assumption that the training set is evenly distributed among classes, this is rarely true in practice. As shown in [11], unbalanced data sets often appear in real world problems. Informative sampling is designed to rapidly model data sets regardless of their label distribution.

Feature selection is a well-studied[5][21][31] field in machine learning. Generally, features are ranked based on their relative influence on outcome. A popular approach is to compare the performance of a certain algorithm by supplying only one feature at a time or steadily enlarging subsets of features. Experiments here show that informative sampling will seek data points with important features automatically without requiring a separate feature selection process.

The following section introduces the general EEA algorithm and the actual algorithm used in this paper. Section 3 first shows the comparative results between informative sampling and competing algorithms, and then the data points that have been chosen by informative sampling are studied. Section 4 concludes the paper.

2. METHODS

In this section, the EEA is introduced, the random and balanced sampling methods are described, and finally the details of the informative sampling algorithm are given.

2.1 Estimation-Exploration Algorithm

The EEA is a coevolutionary algorithm with two stages: the estimation phase and the exploration phase (see Table 1). Each of the two phases has an associated evolutionary algorithm.

Table 1: An outline of the general Estimation-Exploration Algorithm

```
1 Initialization
    1) Create an initial population of candidate models.
    2) Create an initial population of candidate tests
       (virtual data points).
2 Exploration Phase
    1) Evolve candidate tests.
    2) Fitness of a test is the level of disagreement it
       causes among models.
    2) Fitness of a model is its performance on current
       training set.
3 Estimation Phase
    1) Evolve candidate models on the current training
       set.
    3) Data point that causes the most disagreement is
       added into training set.
4 Termination
    Repeat steps 2 and 3 until the population of models
    achieves satisfying performace.
```

The algorithm starts by initializing a population of candidate models and a population of candidate tests. A population of candidate models and a population of candidate tests are randomly generated.

The first pass of the exploration phase begins by supplying the generated candidate tests to the candidate models. The fitness of one data point is the degree of disagreement among the predictions of the models. For a two-class problem, the best possible data point will cause half of the models to predict that it is in one class and the other half to predict it is in another. After a pre-specified period of test evolution, the fittest data point is added into the training set to be used in the estimation phase.

In the estimation phase, the candidate models are then evolved on the training set for a certain number of generations. In each generation, the models with lower fitness are eliminated, and the models with higher fitness are copied and mutated to fill the population. The fitness of a model is its performance on the current training set.

After the completion of the first pass through the estimation phase, the exploration phase is run again. A group of randomly generated candidate tests is supplied to the current candidate models, the tests are re-evolved, and the most fit is added to the training set. The second pass through the estimation phase will then evolve the current models on the updated training set.

Each round of the EEA contains one exploration phase and one estimation phase. The algorithm continues several rounds until certain performance criteria have been satisfied.

The two phases of the EEA can be described as the interactions between two components of the algorithm: the candidate models and the training set. The exploration phase is the process of updating the training set based on the state of the current candidate models. The estimation phase is the process of evolving the candidate models against the training set updated by the exploration phase.

2.2 Random Sampling

Random sampling, in which each data point has the same probability to be chosen, is a popular choice for processing large data sets. It has been widely used in many fields [29][23][26]. In this paper, random sampling is compared against informative sampling.

2.3 Balanced Sampling

Two common methods for processing the unbalanced training sets are over-sampling the minority classes[17] and downsizing the majority class[14], each of which aims to select the same number of data points from each class. This approach is referred to as balanced sampling in this paper. Studies [11] have shown that balanced sampling can help improve the classification performance on certain unbalanced data sets, and it has been used in many applications [19][18]. Though it is a widely accepted method, there is no hard evidence that balanced sampling is the optimal approach. Balanced sampling is here compared against random sampling and informative sampling.

2.4 Informative Sampling

Classifiers used in this paper are artificial neural networks (ANN). The ANN is a standard feed-forward neural network with three layers (input, hidden, output). The input layer has 16 nodes, the hidden layer is set to have 8 nodes and the output layer has one binary node (1 for fatal, and 0 for non-fatal patient outcome). An illustration of the structure of the ANN is given in Figure 1. For each link between nodes, there is a weight associated with it. For each node in the input layer, its value is its corresponding feature value from the current data point. For each node in the hidden layer, its value is calculated as the sum of each weight linked to it multiplied by the weight's corresponding input node. The resulting sum is passed through an activation function. For each node in the output layer, its value is calculated like the nodes in the hidden layer, only now the input values are the values from the hidden nodes. The behavior of the ANN can be summarized by the following function:

$$o = h(\sum_j w_{jk} \times h(\sum_i w_{ij} \times f_i + \theta_j) + \theta_k) \qquad (1)$$

where o is the value of a node in the output layer, i denotes a node in the input layer, j denotes a node in the hidden layer, k denotes the node in the output layer, w_{ij} is the weight from node i to j, w_{jk} is the weight from node j to node k, f_i is the value of the i_{th} feature of an input data point, θj is a constant called the bias of node j, θ_k is the bias of node k, and h is the sigmoid activation function. The value of the output node is a real value between 0 and 1. In this paper, when $o <= 0.5$, the output of the ANN is set to 0; otherwise, the output is 1. In this paper, an evolutionary algorithm(EA) is used to train the ANN. [32] showed that the combination of ANNs and EAs can lead to better intelligent systems than any of them alone.

Table 2: An outline of the informative sampling algorithm

1 Initialization

Randomly create a population of ANNs.

2 Exploration Phase

1) Pass a portion of the data set in as candidate tests.

2) Fitness of a test is the level of disagreement it causes among models.

3) Data point that causes the most disagreement is added into training set.

3 Estimation Phase

1) Apply mutation operator onto each candidate models, an old model is replaced if its child has a better fitness.

2) Fitness of a model is its performance on current training set.

4 Termination

Repeat steps 2 and 3 until the population of models achieves satisfying performace.

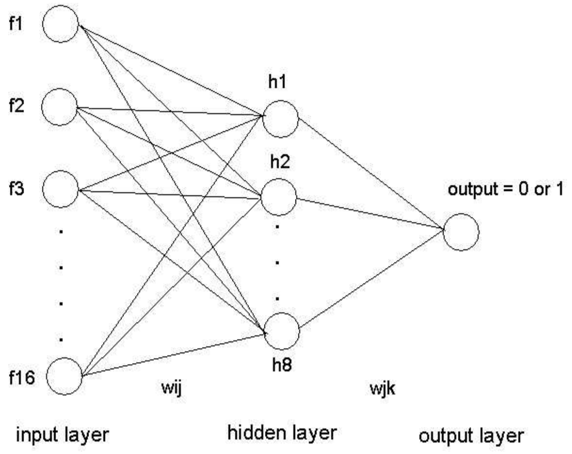

Figure 1: The illustration of the neural network.

The algorithm starts by creating a population of 30 ANNs. The weights of the neural networks are randomly selected real numbers between -1 and 1. The data set is divided into subsets of size p.

The exploration phase is run before the estimation phase at the outset of the algorithm. The first data point to the p_{th} data point are supplied to the current set of candidate models (the random 30 ANNs) sequentially. The predictions of the 30 models on one data point are recorded, and the difference between the number of non-fatal predictions and fatal predictions is then calculated for that point. The data point with the lowest such value is chosen and added into the training set with its corresponding label.

The exploration phase is followed by the estimation phase. At the outset of the first pass through the estimation phase, the random ANNs are trained on the single training datum using:

$$fitness = \frac{c_{nf}}{t_{nf}} \times \frac{c_f}{t_f} \qquad (2)$$

where c_{nf} is the number of correct predictions of non-fatal patients, t_{nf} is the total number of non-fatal patients, c_f is the number of correct predictions of fatal patients, and t_f is the total number of fatal patients. The training is done by applying a mutation operator on each ANN. The fitness of the parent ANN and the child ANN is calculated. If the child ANN is better, the parent model is replaced with it. Otherwise, the parent is retained. One generation is evolved each time the estimation phase is run. This process instantiates a parallel hill climbing algorithm[25]. The mutation operator takes a random weight of the ANN and changes it to a random real value between -1 and 1. The fitness function is suggested specifically for highly unbalanced data sets. A simple fitness function would mislead the models to only output the majority label, regardless of the input. The fitness function used here shapes the ANNs to produce correct predictions on data from both the majority and minority classes.

The algorithm then returns to the exploration phase. This time, the $(p + 1)_{th}$ data point to the $2p_{th}$ data point are supplied to the models that were evolved in the estimation phase. One data point is selected based on model disagreement and added into the training set (which now contains two data points). The estimation phase is run again after the exploration phase on the updated training set. One round of the algorithm consists of a single run of the exploration phase followed by a single run of the estimation phase. The algorithm executes several rounds until certain preset criteria are met.

Several changes are made to the general EEA in order to apply it as a selective sampling method for unbalanced data sets. In previous applications [4][2][3], virtual data points are created as candidate tests. This is legitimate for systems for which training data can be generated on the fly, but not for existing data sets. The solution to this is to use a portion of the data set without their labels as candidate tests.

Table 2 is an outline of the informative sampling algorithm.

3. RESULTS

The informative sampling algorithm was compared against random sampling and balanced sampling. Two sets of comparisons are done: the performance comparison between the informative sampling and the other two sampling algorithms, and a comparison between training data selected by informative sampling and random sampling.

3.1 Characterization of the Data Set

As of January 2007, NTDB has over 1 million records. Each record corresponds to an individual patient. Records with missing data are removed from consideration. The remaining records may serve as training or testing data.

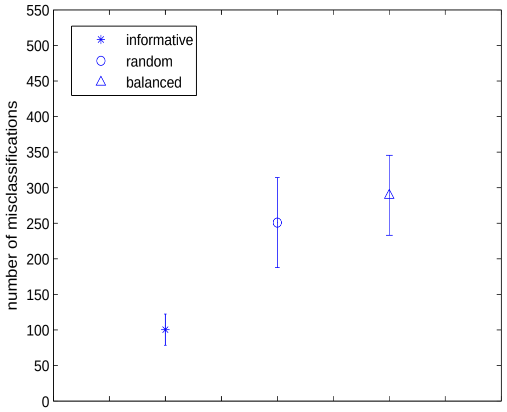

Figure 2: The number of misclassifications comparison among the three algorithms on the unfiltered data set. Error bars indicate two units of standard deviation.

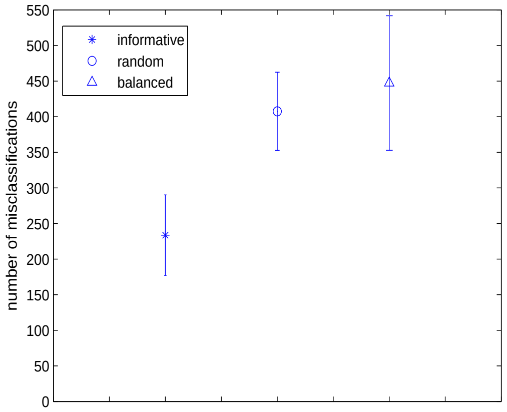

Figure 3: The number of misclassifications comparison among the three algorithms on the filtered data set.

In this paper, classes correspond to discharge status (disstatus). There are two possible classes: fatal(0) or non-fatal(1). The features used are: f1 = age, f2 = gender, f3 = Glasgow Eye Component in ED(edeye), f4 = Glasgow Verbal Component in ED (edverbal), f5 = Glasgow Verbal Component in ED (edmotor), f6 = Glasgow Coma Scale Total in ED (edgcstotal), f7 = blood pressure (fsbp), f8 = Injury Severity Score (iss), f9 = Revised Trauma Score in ED (edrts), f10 = TRISS Survival Probability (probofsurf), f11 = Recalculated edrts by ACS (acs_edrts), f12 = Recalculated probofsurf by ACS (acs_ps), f13 = Glasgow Eye Component at the Scene (sceneeye), f14 = Glasgow Verbal Component at Scene (scenevrb), f15 = Glasgow Motor Component at the Scene (scenemotor) and f16 = Glasgow Coma Scale Total at the Scene (scenegcsto). Each of the features is normalized to a real number between 0 and 1 by divid-

Table 3: The range of features.

Feature	Input Range
age	0-99
gender	Male or Female
edeye	1-4(integer)
edverbal	1-5(integer)
edmotor	1-6(integer)
edgcstotal	1-15(integer)
fsbp	0-300(integer)
iss	0-75(integer)
edrts	0-8(real)
probofsurf	0-1(real)
acs_edrts	0-8(real)
acs_ps	0-1(real)
sceneeye	1-4(integer)
scenevrb	1-5(integer)
scenemotor	1-6(integer)
scenegcsto	1-15(integer)

ing each value by the maximum value for that feature found in the data set. Table 3 outlines the actual range of each feature.

Experiments are conducted on both the whole NTDB (unfiltered) and the NTDB filtered by "headct = positive" (filtered). The unfiltered data set contains all patients that had trauma, while the filtered data set only contains those patients who suffered head trauma. NTDB is highly unbalanced in terms of discharge status. For the unfiltered data set, the fatal to non-fatal ratio is about 1:20, and the ratio for the filtered data set is about 1:6.

3.2 Performance Comparison

In order to fairly compare random and balanced sampling with informative sampling, all three algorithms are executed in the same manner, with the exception of the exploration phase. For random sampling, in the exploration phase, instead of choosing the data point causing the most disagreement, a data point is chosen randomly out of the p data points currently being considered. For balanced sampling, in the exploration phase, data points with fatal and non-fatal outcomes are randomly selected alternatively in each round, which ensures that data points are evenly distributed.

The rest of the settings of the three algorithms remain the same. 30 ANNs are used as the model set. 2000 data points are randomly selected from both the unfiltered and filtered data set as the testing sets. All algorithms run for 600 rounds. Each round consists of one estimation phase and one exploration phase. In each pass through estimation phase, mutation is done once. In each exploration phase, for the unfiltered data set, $p = 100$ data points are fed in sequentially, and 1 is chosen into the training set, for the filtered data set, $p = 30$ data points are supplied sequentially, and 1 is chosen.

Figure 2 reports the performance comparison of the three algorithms in terms of misclassifications on the testing set using the unfiltered data set. Each algorithm is run 30 times independently. It is shown that informative sampling has a mean of 100.4 misclassifications out of 2000 testing data points, compared to random sampling with a mean of 250.9 misclassifications and balanced sampling with a mean of 289.3 misclassifications. As can be seen, informative sampling achieves a marginally smaller standard deviation compared to random and balanced sampling. In this particular case, random sampling performs slightly better than sampling evenly from each class. This indicates that selecting equal numbers of data points from each class does not always confer an advantage.

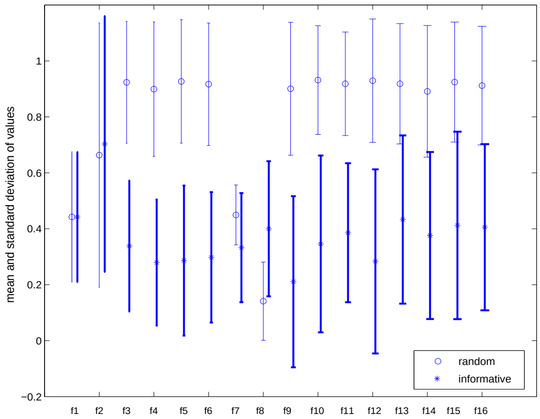

Figure 4: The chosen points comparison of the informative sampling and the random sampling on the unfiltered set. The detailed feature information can be found in section 3.1.

In Figure 3, each algorithm is run 30 times on the filtered data set, which shows a similar pattern to the unfiltered case. Informative sampling has a mean of 233.6 misclassifications, random sampling has a mean of 407.5 misclassifications, and unbalanced sampling has a mean of 447.2 misclassifications. For standard deviation, informative sampling and random sampling are similar, while balanced sampling is noticeably worse due to its larger error bar. Random sampling performs slightly better than balanced sampling in terms of both mean and standard deviation.

3.3 Training Data Comparison

In the previous section, experiments show that informative sampling achieves a better performance in terms of misclassifications on the same amount of training data using the same training method. This indicates that informative sampling obtains a different set of data points for improving classification. In this section, comparison is done between the set of points chosen by informative sampling and random sampling. The chosen points are taken from the experiments described in the previous section on the unfiltered data set using 16 features. There are a total of 30 independent runs. In each run, 600 data points are chosen, giving a total of 18000 data points. The values of the data points are normalized into real numbers between 0 and 1.

Figure 4 reports the means and standard deviations of the values of the chosen points. It is shown that the two algorithms focus on different data points. For "age", there is no difference. For "gender", there is no significant difference. For "edeye, edverbal, edmotor, edgcstotal, edrts, probofsurf, acs_edrts, acs_ps, sceneeye, scenevrb, scenemotor and scenegcsto", informative sampling selects patients with sig-

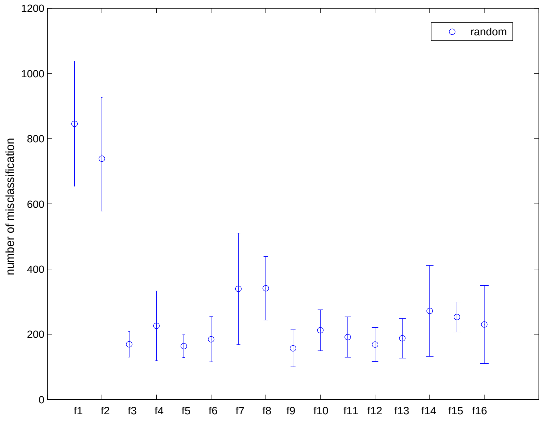

Figure 5: The performance comparison on the random sampling algorithm among individual features. The detailed feature information can be found in section 3.1.

nificant lower values than random sampling. For “fsbp”, informative sampling selects patients with lower blood pressures. For “iss”, informative sampling selects patients with lower severity scores. Note that the data points chosen by the random sampling method have a similar distribution to the entire data set.

Feature selection was then performed on the data set. Each of the 16 features used is provided to the random sampling algorithm as the only input for 30 independent runs. This approach serves as the independent validation for the importance of each feature.

Figure 5 reports the mean performance of random sampling using each of the 16 features only. Again, the performance is the number of misclassifications of the best model of each round. It is shown that “age” is the least informative feature with 823.1 mean misclassifications; “gender” is only marginally better than “age”. All other features are relatively informative, among which “edeye” ,“edmotor” and “edrts” are the most informative.

In Figure 4, informative sampling was shown to select data points that have significantly different values from those selected by the random sampling algorithm on all features except “age” and “gender”. According to Figure 5, the features that informative sampling selects are exactly the features that are informative. Those features were identified automatically by informative sampling, without requiring an additional feature selection process. Therefore while selecting the informative data points, informative sampling is also performing feature selection.

4. CONCLUSIONS AND FUTURE WORK

In this paper, a modified version of the EEA is applied to a large unbalanced medical data set as a selective sampling algorithm, which is referred to as informative sampling. The algorithm uses a set of artificial neural networks as candidate models, and unlabeled samples from the data set as candidate tests. Through the interaction of these two components, the algorithm selects data points that cause the most disagreement among models. The algorithm is applicable to large data sets because it only needs one scan on the whole data set, yet training is carried out only on the chosen training set. Experiments show that the algorithm outperforms the random sampling and balanced sampling, which is a widely used method when dealing with unbalanced data sets. Evidence is shown that the algorithm can automatically select data points with important feature values without requiring a dedicated feature selection process.

ANNs have shown significant promise in modeling outcomes in different types of cancer [10][27], gastrointestinal bleeding [7], and multi-system trauma [8]. The use of ANNs to predict outcome of head trauma was limited to a single model, with a small sample size [9]. This paper has suggested an approach for modeling on a large trauma data set. Although the informative sampling algorithm in this paper uses ANNs as the classification technique, it is a general algorithm that can work with any classifier.

Future prospects of this work will include extending the algorithm by making it determine algorithm parameters (such as the structure of the neural network) automatically, generalizing this algorithm by letting it work with other classifiers and applying the algorithm to various other real world data sets. In addition, because informative sampling takes portions of the data set in for processing in a sequential manner, it will also be applied to streaming data.

5. REFERENCES

[1] J. Bongard and H. Lipson. Automating genetic network inference with minimal physical experimentation using coevolution. In *Proceedings of the 2004 Genetic and Evolutionary Computation Conference*, pages 333–345. Springer, 2004.

[2] J. Bongard and H. Lipson. Active coevolutionary learning of deterministic finite automata. *Journal of Machine Learning Research*, 6:1651–1678, 2005.

[3] J. Bongard and H. Lipson. Nonlinear system identification using coevolution of models and tests. *IEEE Transactions on Evolutionary Computation*, 9:361–384, 2005.

[4] J. Bongard, V. Zykov, and H. Lipson. Resilient machines through continuous self-modeling. *Science*, 314:1118–1121, 2006.

[5] E. F. Combarro, E. Montanes, I. Diaz, J. Ranilla, and R. Mones. Introducing a family of linear measures for feature selection in text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:1223–1232, 2005.

[6] N. Cristianini and J. S. Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[7] A. Das, T. Ben-Menachem, G. Cooper, A. Chak, M. Sivak, J. Gonet, and R. Wong. Prediction of outcome in acute lower-gastrointestinal haemorrhage based on an artificial neural network: internal and external validation of a predictive model. *The Lancet*, 362:1261–1266, 2003.

[8] S. M. DiRusso, A. A. Chahine, T. Sullivan, D. Risucci, P. Nealon, S. Cuff, J. Savino, and M. Slim. Development of a model for prediction of survival in pediatric trauma patients: comparison of artificial

neural networks and logistic regression. *Journal of Pediatric Surgery*, 37:1098–1104, 2002.

[9] S. M. DiRusso, T. Sullivan, C. Holly, S. N. Cuff, and J. Savino. An artificial neural network as a model for prediction of survival in trauma patients: validation for a regional trauma area. *Journal of Trauma*, 49:212–223, 2000.

[10] T. Hanai, Y. Yatabe, Y. Nakayama, T. Takahashi, H. Honda, T. Mitsudomi, and T. Kobayashi. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer Science*, 94:473–477, 2003.

[11] N. Japkowicz. Learning from imbalanced data sets: a comparison of various strategies. In *Proceedings of learning from Imbalanced Data Sets*, pages 10–15. AAAI Workshop, 2000.

[12] J. Y. Jiang, G. Y. Gao, W. P. Li, M. K. Yu, and C. Zhu. Early indicators of prognosis in 846 cases of severe traumatic brain injury. *Journal of Neurotrauma*, 19:869–875, 2002.

[13] P. Juszczak and R. P. W. Duin. Selective sampling based on the variation in label assignments. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 375–378, 2004.

[14] M. Kubat and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

[15] J. A. Langlois, W. Rutland-Brown, and K. E. Thomas. Traumatic brain injury in the united states: emergency department visits, hospitalizations, and deaths. atlanta (ga). Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, 2004.

[16] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of Research and Development in Information Retrieval*, pages 3–12, 1994.

[17] C. X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 73–79. AAAI Press, 1998.

[18] S. M. Lucas and T. J. Reynolds. Learning deterministic finite automata with a smart state labelling evolutionary algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1063–1074, 2005.

[19] S. Luke, S. Hamahashi, and H. Kitano. Genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1098–1105, 1999.

[20] S. G. Machado, G. D. Murray, and G. M. Teasdale. Evaluation of designs for clinical trials of neuroprotective agents in head injury. *Journal of Neurotrauma*, 16:1131–1138, 1999.

[21] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:302–312, March 2002.

[22] I. Muslea, S. Minton, and C. A. Knoblock. Selective sampling with redundant views. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 621–626, 2000.

[23] F. Olken and D. Rotem. Simple random sampling from relational databases. In *12th International Conference on Very Large Data Bases*, pages 160–169, August 1986.

[24] J. R. Quinlan. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[25] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd ed.).* Prentice Hall, Upper Saddle River NJ.

[26] S. L. Salas and E. Hille. *Sampling techniques.* John Wiley and Sons, New York, 1963.

[27] F. Sato, Y. Shimada, F. M. Selaru, D. Shibata, M. Maeda, G. Watanabe, Y. Mori, S. A. Stass, and M. I. S. J. Meltzer. Prediction of survival in patients with esophageal carcinoma using artificial neural networks. *Cancer*, 103:1596–1605, 2005.

[28] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pages 287–294, 1992.

[29] J. Waksberg. Sampling methods for random digit dialing. *American Statistical Association*, 73(361):40–46, March 1978.

[30] P. D. Wasserman. *Advanced Methods in Neural Computing, 1st edition.* John Wiley and Sons, Inc., New York, NY, USA.

[31] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, 1997.

[32] X. Yao. Evolving artificial neural networks. In *Proceedings of the IEEE*, pages 1423–1447, 1999.