Published in final edited form as:

Neural Netw. 2012 January ; 25(1): . doi:10.1016/j.neunet.2011.07.002.

The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support

Jordan M. Malof^{a,*}, Maciej A. Mazurowski^b, and Georgia D. Tourassi^{b,2}

^aDepartment of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40208, USA

^bCarl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC 27705, USA

Abstract

Case selection is a useful approach for increasing the efficiency and performance of case-based classifiers. Multiple techniques have been designed to perform case selection. This paper empirically investigates how class imbalance in the available set of training cases can impact the performance of the resulting classifier as well as properties of the selected set. In this study, the experiments are performed using a dataset for the problem of detecting breast masses in screening mammograms. The classification problem was binary and we used a k -nearest neighbor classifier. The classifier's performance was evaluated using the Receiver Operating Characteristic (ROC) area under the curve (AUC) measure. The experimental results indicate that although class imbalance reduces the performance of the derived classifier and the effectiveness of selection at improving overall classifier performance, case selection can still be beneficial, regardless of the level of class imbalance.

Keywords

Case-based reasoning; Computer-aided decision; Class imbalance; Case selection

1. Main text

In machine learning, the binary classification problem is well known and has garnered much attention in the literature (Bishop, 2006; Duda, Hart, & Stork, 2001). A large variety of classifiers have been proposed over the years, creating a diverse array of choices for a given classification problem. Despite their variety, most classifiers follow one of the two fundamental lines of reasoning: either rule based reasoning or case-based reasoning. Rule based classifiers such as neural networks (Zurada, 1992) or decision trees (Bishop, 2006) operate by using available training cases (also called instances Aha, Kibler, & Albert, 1991) to extract rules with which future query cases can be classified. In case-based reasoning, new query cases are classified by directly comparing them to template cases from each class. These template cases are maintained in a database called the case base.

One major advantage of case-based systems over rule based systems is that they do not use any rules/abstractions that must be updated each time the system is extended to utilize more example (or training) cases. Neural networks for example must be retrained if more cases are to be incorporated into their connection weights. In case-based systems, utilizing a new case simply requires that it is added to the already existing case base. However, indefinitely storing new cases in a database has certain limitations.

Generally, when cases are randomly sampled from a population, more samples will provide better representation (Raudys & Jain, 1991). Therefore, maintaining more cases in the case base should result in better classification performance. Unfortunately, as more cases are collected and placed in the case base, several practical problems can arise. One such problem is storage limitation. While the cases stored in many case-based systems are feature vectors, a relatively compact representation, some systems use entire images (Tourassi, Vargas-Voracek, Catarious, & Floyd, 2003). In such systems, any growth in the case base, while potentially improving performance, can also make the case base impractically large.

Another problem occurs when the computation time required for classification becomes excessive. Since a case-based classifier must compare the query case to every stored case, the computation cost of performing a classification task grows in proportion to the size of the case base. In practical systems the computational power is limited and therefore, as the case base grows larger so does the time required for classification. If the case base grows too large, the computational time can become impractical. This is especially true for systems where assessment of similarity (or distance) between cases is associated with a large computational cost (Tourassi et al., 2003).

1.1. Case selection

Each disadvantage that was mentioned above results directly from excessive case base size. Therefore, by reducing the size of the case base these problems can be reduced or even eliminated. Obviously, if cases are removed from the case base there may be an associated loss in classification accuracy. This presents a new problem: how to remove cases in such a way that it minimizes any potential loss in classifier performance? While randomly selecting a subset of cases is possible, a more sophisticated approach can be beneficial. Multiple case selection algorithms have been proposed for this task (Wilson & Martinez, 2000). Many of these algorithms have been shown to significantly outperform random selection (Mazurowski, Malof, & Tourassi, 2011). Further, some algorithms such as Random Mutation Hill Climbing have even been shown to improve classifier performance, while providing significant reduction in the size of the case base (Mazurowski, Malof, & Tourassi, 2011).

While case selection algorithms have been shown to be effective, their performance may be dependent on the characteristics of the available data. This is readily observable in Wilson and Martinez (2000), where each selection algorithm had varying impact on classification performance depending on the dataset that was used. Therefore, depending on the data, a given selection algorithm can have a very different impact on performance.

This dependency on data is better understood given a few basic considerations. Given a set of cases representing the distributions of two classes, cases can be broadly described in one of the following ways:

- Redundant
- Misleading
- Beneficial.

Redundant cases are very similar or the same as one or more cases in the same class. Therefore removing a redundant case may not be detrimental to classification, while allowing for the benefits of shrinking the case base.

Misleading cases are those cases that may actually be detrimental to classifier performance, either because they are noisy or they are outliers. Removing cases of this type has the potential to be beneficial to classifier performance while also allowing for reduction of the case base.

Finally, beneficial cases are those cases that are generally useful for classification. These cases are distinguishable because they do not fall into the other categories. While redundant and misleading cases can be beneficial for classification, in general their removal is not detrimental to classification and it allows for case base reduction. If a dataset is composed solely of beneficial cases, case selection is guaranteed to have a detrimental effect on classifier performance.

The above considerations help highlight the inherent sensitivity of case selection to the data upon which it is performed. Different applications involve different data and therefore the effectiveness of case selection depends on the application for which it is used and the properties of the corresponding data.

1.2. Class imbalance

Multiple properties of the training data can affect performance of the case selection algorithm. One of these properties is presence of class imbalance. Class imbalance occurs when one of the classes is underrepresented. This occurrence is often a result of data collection limitations. Specifically, it may occur in computer-aided decision (CAD) systems. Such systems have become popular as diagnostic aids in numerous branches of medicine such as screening mammography (Cheng et al., 2006; Tourassi, Lo, & Floyd, 2007) and chest imaging (Wu, Doi, Kobayashi, MachMahon, & Giger, 1997). The data used to train such systems is collected from real patient populations. When the prevalence of the disease of interest is very low, the “healthy” class is far better represented than the “diseased” class in the collected data. Since case selection is sensitive to the training data, it is important to understand the impact such class imbalances might have on selection and the resulting case-based classifiers.

In this paper, we perform an empirical study to gain insight into the impact of class imbalance on case selection and the resulting classifier. We perform our study using data for the problem of breast cancer detection in mammograms.

The remainder of the paper is organized as follows. Section 2 describes the materials and methods that were used in our experiments. This includes a description of the selection algorithm and the case-based classifier. Section 3 describes the experimental design. This section includes the dataset used in the study, the experimentation parameters, and the data handling scheme. Section 4 outlines the results obtained from the experiments and Section 5 is a discussion of significant results and their implications.

2. Methods

For the remainder of this paper, for a given classification problem, T will be used to denote the set of training cases. The selected subset that results from case selection will be denoted by S . This study is limited to a single case-based classifier and a single selection algorithm for experimentation. Both are described below.

2.1. The case-based classifier: k-nearest neighbor algorithm

K-Nearest Neighbor (*k*-NN) classifier is well known and has been used in numerous case-based classification studies (Duda et al., 2001; Horton & Nakai, 1997; Wilson & Martinez, 2000). *K*-NN was selected for use as the case-based classifier in this study primarily because of its broad use and understanding in literature. The algorithm proceeds as follows. Given a query case, its distance (or dissimilarity) to all the cases stored in the case base is calculated. Then *k* cases with the lowest distance to the query case are chosen. Based on these chosen cases a decision index is calculated.

$$f(x_q) = \frac{\sum_{i=1}^k \omega_i v(x_i)}{\sum_{i=1}^k \omega_i} \quad (1)$$

where

$$\omega_i = \frac{1}{d(x_q, x_i)^2} \quad (2)$$

Here, x_q is a query case being presented for classification. The value $d(x_q, x_i)$ is its distance to each template case x_i in the case base. The function $v(.)$ maps each template case to its respective class membership (i.e., 1 or 0). The function $f(.)$ then returns the decision index of the query case. To determine the estimated class membership of the query case, a threshold $\mathbf{t}$ is applied such that, if the decision index is higher than $\mathbf{t}$ the query case is classified as positive, and if the decision index is lower than $\mathbf{t}$ the query case is classified as negative.

In this study we used Euclidean distance in feature space to assess dissimilarity between cases. For a more comprehensive review of the *k*-NN algorithm please see Duda et al. (2001).

2.2. The selection algorithm: random mutation hill climbing

To perform selection in our experiments we used the random mutation hill climbing (RMHC) algorithm (Skalak, 1994). RMHC was chosen due to its good empirical performance in literature (Mazurowski, Malof, Zurada, & Tourassi, 2011; Wilson & Martinez, 2000) as well as its simplicity. In addition, RMHC allows the user to choose how many cases will be included in the final selected subset. This is not true of most case selection algorithms. Thus RMHC gives the designer additional control when designing the system with imposed storage constraints.

The RMHC selection algorithm begins with random selection of a subset $\mathbf{S}$ from the available set of training cases $\mathbf{T}$. The size of the selected subset ($|\mathbf{S}|$) is chosen by the system designer. This value is generally denoted by $\mathbf{m}$. The algorithm then proceeds by switching a randomly selected case in $\mathbf{S}$ with a randomly selected case from the remaining cases ($\mathbf{T} - \mathbf{S}$). The performance of the augmented set $\mathbf{S}$ is then evaluated on the entire training set ($\mathbf{T}$) using the leave one out approach. If the change improves the classification performance it is retained. This process continues for 2000 iterations in an effort to improve the performance of the classifier based on $\mathbf{S}$. The implicit assumption in this approach is that $\mathbf{T}$ is

representative of the true class distributions and therefore, if the classifier based on **S** has good classification accuracy on **T** it will have good performance in general.

3. Experimental design

The primary goal of this study was to assess the impact of class imbalance in training data as it pertains to case selection and the performance of the resulting classifier. Individual experiments consisted of a training set of cases and a test set of cases. The training set here represents the data available to the system designer, and therefore will be subject to controlled changes in prevalence. The test set represents unknown query cases and therefore will have a constant level of prevalence throughout experimentation.

For a single experiment, case selection was performed on the training set cases. The resulting subset was then utilized with the k -NN classifier to classify the test set cases. The level of prevalence in the training data was systematically changed to assess its impact on selection and classifier performance. For a set of training data **T** we will now define the class prevalence **c** below:

$$c = \frac{n_{\text{pos}}}{n_{\text{tot}}}. \quad (3)$$

Here n_{pos} is the number of positive class cases in **T** and n_{tot} is the total number of cases in **T**. More generally, the **c** index provides a measure of the prevalence on any given set of cases.

The impact of **c** was assessed by measuring the prevalence in the selected set **S** for each training set **T**. The impact of prevalence on classification was assessed by measuring the performance of each derived k -NN classifier on the test set. Performance was measured using the receiver operator characteristic (ROC) area under the curve (AUC) measure (Bradley, 1997). This is a commonly used measurement in medical CAD systems (Mazurowski, Malof, & Tourassi, 2011; Obuchowski, 2003). The k -NN algorithm for classification operated using $k = 20$.

3.1. Dataset description

The approached problem was distinguishing between masses (which may be indication of breast cancer) and normal tissue in mammograms. We extracted 1800 regions of interest (ROIs) from mammograms included in Digital Database of Screening Mammography (DDSM) (Heath et al., 1998). Each ROI was 512×512 pixels in size. 900 ROIs depicted biopsy-proven masses (positives) and 900 depicted normal tissue (negatives). From each ROI, seven features were extracted: mean intensity, standard deviation of the intensity, contrast, correlation, energy, homogeneity, and entropy. There was no feature selection and the features were normalized so that each feature had a mean of 0 and a standard deviation of 1.

3.2. Data handling

To properly assess performance of classifiers, two disjoint sets were constructed from the 1800 available cases: a test set and a training set. The test set was constructed by randomly drawing 200 cases from the 1800. A single constraint was placed on this random drawing: the resulting test set had to have equal class representation (**c** = 0.5).

Once a test set was created, six training datasets were randomly drawn from the remaining 1600 cases. Only 800 cases were drawn for each training set. This size was chosen in order to facilitate testing for all the desired levels of class imbalance while maintaining consistent

training set size. Each training set was drawn such that it had a different level of class prevalence c . Therefore, there were six different levels of prevalence used in this study: $c = (0.01, 0.02, 0.05, 0.1, 0.2, 0.5)$. Then, for each training set the RHMC algorithm was executed. Eleven different values of parameter m (desired case base size) were used ranging from $m = 20$ to $m = 800$. Therefore for each training set, eleven runs of RMHC were executed. Please note that at $m = 800$ there is no case selection because this is the size of the training data.

Then for each training dataset (with different extents of class imbalance) and selected set that was generated (different desired case base sizes), the performance of the resulting case-based classifier was evaluated using the test set.

The entire process described above was repeated 40 times, each time with a different randomly drawn train/test set.

4. Experimental results

The data analysis was focused on the impact of class imbalance to: (i) classifier performance, (ii) class prevalence in the resulting set of selected cases. Results are presented accordingly.

4.1. Impact on classifier performance

The results of the experiments on the impact of class imbalance on classifier performance are summarized in Fig. 1. The plot in Fig. 1 has 6 different curves corresponding to the 6 different levels of prevalence that existed in the original training sets. In each of these curves, average AUC test performance is depicted as a function of changing the selected set size m parameter. All 6 curves illustrate the same general trend. The performance of the classifier decreases as the desired size m of the selected case base increases. The performance of the classifiers appears to improve again as m gets closer to the original case base size at $m = 800$.

A few notable observations can be made about these results. First, for any given value of m , classifier performance always deteriorates as c decreases. In other words, class imbalance always damages classifier performance, independent of case selection. This corroborates our previously reported results (Malof, Mazurowski, & Tourassi, 2009). Second, as class imbalance increases (i.e. c decreases), the selected set size m has a higher impact on the performance of the classifier resulting from selection. It is interesting to note, however, that there is always the potential for selection with RMHC to improve the classifiers expected performance, particularly at lower levels of prevalence c .

Finally, observing the curve at $c = 0.5$, performance actually improves as the case base size is reduced. This result is consistent with results obtained in other studies involving RMHC (Mazurowski, Malof, Zurada, et al., 2011; Wilson & Martinez, 2000).

4.2. Impact on the class prevalence of the selected set

The impact of imbalance on case selection was assessed by measuring the class prevalence in the case base after selection was performed. These results are summarized in Fig. 2. Similar to Fig. 1 there are 6 separate curves, one for each of the 6 levels of imbalance that were evaluated. Note that the variable c in Fig. 2 denotes the prevalence of the training set upon which selection was performed. All curves in Fig. 2 follow the same general trend. At small selected set sizes ($m < 80$), the resulting prevalence reaches a maximum, being greater than the original case base prevalence. This finding suggests that RMHC tends to select more balanced datasets. An interesting effect occurs as the case base size is increased

however. The prevalence begins to drop until it is below the original prevalence level. When the selection size m becomes close to the original case base size it abruptly rises again until it reaches the original training set prevalence (because no selection is performed at $m = 800$).

A few notable observations can be made based on these results. First, Figs. 1 and 2 together suggest that classifiers constructed with more balanced class representation (e.g. c closer to 0.5) tend to yield higher classification. Second, there appears to be a relationship between the selected set prevalence c and the selected case base size m . It is interesting to note that as m increases, the prevalence of the resulting selected set begins to drop below the original value c in the training set. This same trend occurs for each curve except for the scenario when the two classes are equally represented ($c = 0.5$). This result is consistent with results obtained in our previous work with artificial data (Malof et al., 2009).

5. Conclusions

In this study we examined the impact of class imbalance within the training data used for case selection and classification for a case-based classifier. This work was an extension of earlier work where similar experiments were performed on artificially generated data (Malof et al., 2009). The conclusions of the original study are validated by the results obtained in this study using real clinical data.

Several general conclusions can be drawn from these studies:

- Class imbalance in the training data is always detrimental to classification performance, independent of case selection.
- For any level of class imbalance c there always exist one or more choices for the selected set size m that will yield better classification performance than the classifier based on the entire case-base.
- Conversely, for any level of class imbalance, there exist one or more choices for the selected set size m that will yield worse classification performance than the classifier based on the entire case-base.
- The higher is the class imbalance in the case base, the stronger is the beneficial effect of selection on classification performance.

It is important to note that while these conclusions hold for our particular experimental design (i.e. Selection with RMHC, classification with k -NN, etc.) we have not presented analytical proof that these conclusions can be asserted for all case selection problems with class imbalance. We believe, however, that the results and conclusions presented are likely to hold across a broad number of scenarios, or otherwise offer useful insight into them. We make this assertion based on the relatively representative nature of k -NN and RMHC. k -NN is very popular and extensively used as a case-based classifier in the machine learning literature. RMHC is also popular as a selection algorithm (Cano, Herrera, & Lozano, 2003; Jankowski & Grochowski, 2004; Wilson & Martinez, 2000). Additionally, RMHC operates by optimizing the selected cases using their classification performance on the training set. This principle is shared by many case selection algorithms, often the only difference being the addition of one or more heuristics. Some notable examples of such related algorithms include the CNN (Hart, 1968), RNN (Gates, 1972), MC-1 (Skalak, 1994), Explore (Cameron-Jones, 1995), DROP1–DROP5 (Wilson & Martinez, 2000), Genetic Algorithms (Cano et al., 2003).

The biggest limitation of this study stems from our use of a single classification problem: we used a real-world dataset for the classification of masses and normal tissue in mammograms.

Despite this, it is important to note that many of the conclusions from this paper are consistent with those obtained from our preliminary study on this topic (Malof et al., 2009). In the preliminary study we used artificially generated data with different characteristics than those of the data presented here (different number of features, different number of cases and likely different probability distributions of the data). This suggests that the conclusions of this study might generally hold for such problems, however, further studies are needed to confirm this. As such, a comparative study of class imbalance utilizing several classification problems would be an important addition to any future work.

In this study we analyzed the impact of imbalanced class representation on case-based classifiers and case selection strategies. Our study confirms that class imbalance has a significant impact on case-based classifiers and case selection. Class imbalance is an important aspect of experimental design and researchers who face such data constraints (particularly in the clinical domain where class imbalance is a frequent occurrence) should be aware of the associated implications.

Acknowledgments

The work of Georgia Tourassi was supported in part by grant R01 CA101911 from the National Cancer Institute.

References

- Aha DW, Kibler DW, Albert MK. Instance-based learning algorithms. *Machine Learning*. 1991; 6(1): 37–66.
- Bishop, CM. *Pattern recognition and machine learning*. New York, NY: Springer; 2006.
- Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 1997; 30(7):1145–1159.
- Cameron-Jones, RM. Instance selection by encoding length heuristic with random mutation hill climbing. *Proceedings of the eighth Australian joint conference on artificial intelligence*; 1995. p. 99–106.
- Cano JR, Herrera F, Lozano M. Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. *IEEE Transactions on Evolutionary Computation*. 2003; 7(6):561–575.
- Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN. Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition*. 2006; 39(4):646–668.
- Duda, RO.; Hart, PE.; Stork, DG. *Pattern classification*. 2. New York, NY: Wiley-Interscience; 2001.
- Gates GW. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*. 1972; 18(3): 431–433.
- Hart PE. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*. 1968; 14:515–516.
- Heath M, Bowyer KW, Kopans D, Kegelmeyer P, Moore P, Chang R, Munishkumaran K. Current status of the digital database for screening mammography. *Computational Imaging and Vision*. 1998; 13:457–460.
- Horton, P.; Nakai, K. Better prediction of protein cellular localization sites with the K nearest neighbors classifier. *Proceedings of the 5th international conference on intelligent systems for molecular biology*; AAAI Press; 1997. p. 147–152.
- Jankowski N, Grochowski M. Comparison of instance selection algorithms II. results and comments. *Lecture Notes in Computer Science*. 2004; 3070:580–585.
- Malof, JM.; Mazurowski, MA.; Tourassi, GD. The effect of class Imbalance on case selection for case-based classifiers, with emphasis on computer-aided diagnosis systems. *Proceedings of the international joint conference on neural networks*; 2009. p. 1975–1980.
- Mazurowski MA, Malof JM, Tourassi GD. Comparative analysis of instance selection algorithms for instance-based classifiers in the context of medical decision support. *Physics in Medicine and Biology*. 2011; 56(2):473–489. [PubMed: 21191152]

- Mazurowski MA, Malof JM, Zurada JM, Tourassi GD. Comparative analysis of instance selection algorithms for instance-based classifiers in the context of medical decision support. *Physics in Medicine and Biology*. 2011; 56(2):473–489. [PubMed: 21191152]
- Obuchowski NA. Recovering operating characteristic curves and their use in radiology. *Radiology*. 2003; 229:3–8. [PubMed: 14519861]
- Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions of Pattern Analysis and Machine Intelligence*. 1991; 13:252–264.
- Skalak, DB. Prototype and feature selection by sampling and random mutation hill climbing algorithms. *Proc. 11th int'l conf. machine learning*; 1994. p. 293-301.
- Tourassi GD, Lo JY, Floyd CE. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. *Medical Physics*. 2007; 34(1)
- Tourassi GD, Vargas-Voracek R, Catarious DM Jr, Floyd CE Jr. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. *Medical Physics*. 2003; 30(8):2123–2130. [PubMed: 12945977]
- Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. *Machine Learning*. 2000; 38(3):257–286.
- Wu XW, Doi K, Kobayashi T, MachMahon H, Giger ML. Development of an improved CAD scheme for automated detection of lung nodules in digital chest images. *Medical Physics*. 1997; 24(9): 1395–1403. [PubMed: 9304567]
- Zurada, JM. *Introduction to artificial neural systems*. St. Paul, MN: West Publishing Co; 1992.

Fig. 1.
Performance of the classifier as a function selected case base size.

Fig. 2.

Prevalence in the selected case base as a function of selected case base size (m).