Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2010 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2010

Supporting Acute Appendicitis Diagnosis: A Pre-Clustering-Based Classification Technique

Yen-Hsien Lee
National Chiayi University, yhlee@mail.ncyu.edu.tw

Paul Jen-Hwa Hu
University of Utah, paul.hu@business.utah.edu

Wei-Yao Chuang
National Chiayi University, s0971416@mail.ncyu.edu.tw

Tsang-Hsiang Cheng
Southern Taiwan University, cts@mail.stut.edu.tw

Yi-Cheng Ku
Providence University, justin.yc.ku@gmail.com

SUPPORTING ACUTE APPENDICITIS DIAGNOSIS: A PRE-CLUSTERING-BASED CLASSIFICATION TECHNIQUE

Yen-Hsien Lee, Department of Management Information Systems, National Chiayi University, Chiayi, Taiwan, R.O.C., yhlee@mail.ncyu.edu.tw

Paul Jen-Hwa Hu, Department of Operations and Information Systems, University of Utah, Salt Lake City, UT, USA, paul.hu@business.utah.edu

Wei-Yao Chuang, Department of Management Information Systems, National Chiayi University, Chiayi, Taiwan, R.O.C., s0971416@mail.ncyu.edu.tw

Tsang-Hsiang Cheng, Department of Business Administration, Southern Taiwan University, Tainan, Taiwan, R.O.C., cts@mail.stut.edu.tw

Yi-Cheng Ku, Department of Computer Science & Information Management, Providence University, Taichung, Taiwan, R.O.C., justin.yc.ku@gmail.com

Abstract

Service quality and cost containment represent two critical challenges in healthcare management. Toward that end, acute appendicitis, a common surgical condition, is important and requires timely, accurate diagnosis. The diverse and atypical symptoms make such diagnoses difficult, thus resulting in increased morbidity and negative appendectomy. While prior research has recognized the use of classification analysis to support acute appendicitis diagnosis, the skewed distribution of the cases pertaining to positive or negative acute appendicitis has significantly constrained the effectiveness of the existing classification techniques. In this study, we develop a pre-clustering-based classification (PCC) technique to address the skewed distribution problem common to acute appendicitis diagnosis. We empirically evaluate the proposed PCC technique with 574 clinical cases of positive and negative acute appendicitis obtained from a tertiary medical center in Taiwan. Our evaluation includes tradition support vector machine, a prevalent resampling classification technique, Alvarado scoring system, and a multi-classifier committee for performance benchmark purposes. Our results show the PCC technique more effective and less biased than the benchmark techniques, without favoring the positive or negative class.

Keywords: Acute Appendicitis, Imbalanced Dataset, Classification Analysis, Re-sampling Approach.

1. INTRODUCTION

The growing demand on service quality and cost containment challenge the healthcare organizations around the world (Kraft et al. 2003; Walczak & Scharf 2000); they are particularly crucial with respect to common illness or medical conditions. For example, acute appendicitis is a common surgical condition that requires timely, accurate diagnosis (Prabhudesai et al. 2008). Delayed or missed diagnoses can increase the morbidity and prolong patients' hospital stay because of serious complications (Flum & Koepsell 2002; Chen & Hu 2009; Ditillo et al. 2006). On the other hand, negative (unnecessary) appendectomies demand significant patient care costs that should be avoided altogether. The average cost of a negative appendectomy was estimated at EUR 2,712 in Europe (Bijnen et al 2003) and the annual expenses of negative appendectomies in the United States amounted to USD 740 million (Flum & Koepsell 2002).

Acute appendicitis diagnosis involves patient history and clinical examination. Diverse and atypical symptoms make timely, accurate diagnosis of acute appendicitis difficult (Birnbaum & Wilson 2000). Typical symptoms of acute appendicitis (such as right lower quadrant pain, abdominal rigidity, and migration pain) are observed with 50% to 67% of the patients only (Old et al. 2005; Filewood 2005). The likelihood of missed diagnoses by experienced clinicians was estimated between 20% and 33% (Old et al. 2005; Pieper et al. 1982), and the probability of negative appendectomy between 9% and 40% (Liang 2005; Prabhudesai et al. 2008; Bijnen et al. 2003; Birnbaum & Wilson 2000). Although an extended observation of the patient could improve the diagnostic accuracy, it likely will cause serious complications because of delaying an appendectomy (Ditillo et al. 2006). The use of imaging technologies, including ultra-sonography (US) and computed tomography (CT), has been shown to improve the diagnostic accuracy; however, the cost and availability make their routine use impractical (Prabhudesai et al. 2008).

The recent advancements of information technology and the growing use of databases have encouraged the data mining approach for reducing the likelihood of misdiagnosis as well as avoiding unnecessary surgeries through systematic analyses of patients' medical records (Kraft et al. 2003; Walczak & Scharf 2000; Delgado et al. 2001). Toward that end, classification analysis is essential, hereby revealing important patterns or rules that separate the clinical cases of positive versus negative diagnosis. A handful of algorithmic techniques support effective classification analysis that constructs a prediction model from a training sample consisted of clinical cases. Common techniques cal be categorized as decision tree induction (e.g., ID3 (Quinlan 1986) and C4.5 (Quinlan 1993)), decision rule induction (e.g., CN2 (Clark & Boswell 1991)), neural network (e.g., backpropagation neural network (Rumelhart et al. 1986)), Bayesian classification (Heckerman 1997), Support Vector Machine (SVM) (Vapnik 1995), or *k*-nearest-neighbor classification (Dasarathy 1991). Previous research has applied such classification techniques to support acute appendicitis diagnosis by analyzing and discovering key diagnostic patterns obscured in the voluminous patient medical records. For example, Zorman et al. (2001) induce a decision tree from the anamneses for predicting acute appendicitis, which exhibits a prediction specificity rate ranging 60% and 83%. Prabhudesal et al. (2008) employ the artificial neural network technique to support acute appendicitis diagnosis and report a reasonably good prediction performance.

Although prior studies recognize the value of classification analysis for acute appendicitis diagnosis, they fail to address the problem associated with a skewed distribution of the examples in the training sample used for building an automated prediction model; that is, a training sample consists of a large number of cases pertaining to one outcome class and only few examples of cases pertinent to the other outcome class (Chawla et al. 2002; Cheng & Hu 2009; Kubat & Matwin 1997). Such skewed distribution problems have been shown to affect the performance of a classification technique significantly (Cheng & Hu 2009; Chan et al. 1999; Japkowicz 2000). Skewed distribution problems exist in various real-world classification scenarios (Chan et al. 1999; Kubat et al. 1998; Cheng & Hu 2009); they are common in health care; e.g., disease predictions involving a positive or a negative diagnosis. Take acute appendicitis for example: we can easily collect a large number of positive cases (i.e., patients diagnosed of having acute appendicitis) but often have difficulty gathering negative

cases from patients who suffer from other illnesses or diseases (i.e., non-acute appendicitis). In effect, negative cases (i.e., non-acute appendicitis) are available through a pathological examination of the negative appendectomy (Ng et al. 2007; Terasawa et al. 2004).

To address the skewed distribution problem in the training sample, prior research has applied such techniques as resampling; e.g., over-sampling, and under-sampling. The use of over-sampling increases the number of the minority cases in the training sample by randomly duplicating or synthesizing the instances pertaining to the minority outcome class (Lewis & Catlett 1994; Chawla et al. 2002). In contrast, under-sampling maintains the integrity of the minority cases by reducing the number of the majority cases in the training sample. In both cases, the intent is to create a training sample consisted of a comparable number of instances between the outcome classes. However, re-sampling is associated with sampling biases that in turn constrain the consistency of its results.

We address the skewed distribution problem inherent to constructing a prediction model for acute appendicitis diagnosis by developing a pre-clustering-based classification technique (PCC), which is more effective than most prevalent classification techniques. The remainder of the paper is organized as follows: In Section 2, we provide an overview of acute appendicitis and diagnosis, and review prior research examining classification analyses involving a skewed distribution of instances in the training sample. In Section 3, we describe the proposed PCC technique, followed evaluation design details and key results in Section 4. We conclude in Section 5 with a summary and discussions of the study's contributions and some important future research directions.

2. LITERATURE REVIEW

In this section, we provide an overview of the acute appendicitis and common diagnostic methods, and review previous research examining classification analysis and imbalanced sample problem.

2.1 Acute Appendicitis and Common Diagnosis Methods

Vermiform appendix is an elongated diverticulum with lymphocytic follicles (Graffeo & Counselman 1996). Obstruction of intraluminal space in appendix can cause the production of mucosa and the proliferation of bacteria, which pressure appendicial wall and thus elicit the clinical symptoms (signs) of appendicitis. Acute appendicitis is a common surgical condition requiring a timely, accurate diagnosis (Prabhudesai et al. 2008). On average, the likelihood that an individual suffers from appendicitis in his or her life span is approximately 7% (Hardin 1999). Acute appendicitis can occur at any age, but more common with people between 10 and 30 years old (Schwartz 1994; Pieper et al. 1982). Although the fatality of acute appendicitis is relatively low (i.e., less than 1%), a delayed or missed diagnosis can lead to appendicial ruptures and other lethal complications.

The most common symptom of acute appendicitis is abdominal pain; other symptoms include anorexia, nausea, vomiting and migratory pain. Abdominal pains are often observed in many other illnesses and diseases, thus making an early, accurate diagnosis of appendicitis challenging (Hardin 1999). To assist acute appendicitis diagnosis by clinicians, several scoring systems are developed. For example, the Alvarado scoring system, also known as MANTRELS, represents a salient scoring system among clinicians (Alvarado 1986; Abdeldaim et al. 2007). It depends on eight predictive factors for acute appendicitis diagnosis, including right lower quadrant pain, leukocytosis, migration of pain to right lower quadrant, anorexia-acetone, nausea or vomiting, rebound pain, elevation of temperature, and the shift of white blood cell count to left. The two most important factors, right lower quadrant pain and leukocytosis, are assigned a score of two each, and the other factors are assigned a score of one each for a total score of 10. A patient should be considered as suffering from acute appendicitis if the total score reaches 7 or higher. The use of imaging technologies, such as ultrasonography and computed tomography (CT) also prevails and has been shown valuable to acute appendicitis diagnosis (Al-Khayal & Al-Omran 2007; Mun et al. 2006). Although ultrasonography can improve acute appendicitis diagnosis, as suggested by its achieving a high sensitivity rate, the actual diagnostic value (accuracy) seems to vary greatly among operators who often differ significantly in skill levels and clinical experiences (Old et al. 2005). CT is generally more accurate

than ultrasonography but its use involves radiation exposures, substantially higher costs, and limited availability. Furthermore, the toxicity of contrast media may create negative impacts on patients or even cause other complications (Old et al. 2005). Together, these considerations confine the use of CT as a regular mean for acute appendicitis diagnosis clinically.

2.2 Classification Analysis and Imbalanced Dataset Learning

Classification analysis entails a process of automatically constructing a classification model on the basis of important relationships between the decision classes (i.e., outcomes) and the values of specified input attributes (i.e., decision variables) in a sample of preclassified examples (i.e., training instances). Such a classification model, once induced, then can be used to classify new instances. A host of different algorithmic classification techniques have been developed, which can be broadly categorized as decision tree induction (e.g., ID3 (Quinlan 1986) and C4.5 (Quinlan 1993)), decision rule induction (e.g., CN2 (Clark & Boswell 1991)), neural network (e.g., backpropagation neural network (Rumelhart et al. 1986)), Bayesian classification (Heckerman 1997), support vector machine (SVM) (Vapnik 1995), or *k*-nearest-neighbor classification (Dasarathy 1991).

An imbalanced sample has a skewed distribution of preclassified examples across different outcome classes. While most existing classification techniques assumes a balanced training sample, many real-world classification scenarios are characterized by a skewed distribution of instances pertaining to different outcome classes; e.g., fraud detection (Chan et al. 1999), oil spill detection (Kubat et al. 1998), healthcare management (Cheng & Hu 2009), disease prediction. A review of prior research results shows that a classification model induced from the imbalanced dataset have a high tendency of classifying a new instance to be of the major (outcome) class (Cheng & Hu 2009; Chan et al. 1999).

Re-sampling, over-sampling or under-sampling, has been applied to address the imbalanced training sample problem. In essence, over-sampling approach increases the number of examples pertaining to the minority class through random duplications or synthetic instance creations. Lewis and Catlett (2002) duplicate the instances of the minority class in a training sample in order to balance against those of the majority class. DeRouin et al. (1991) use an artificial neural network to synthesize the training instances. Chawla et al. (2002) propose a technique for synthesizing learning instances, synthetic minority over-sampling technique (SMOTE), which synthesizes new instances by randomly generating a data point between the randomly selected instance of the minority class and its nearest neighbor. Specifically, SMOTE randomly selects an instance from the minority class and then searches for its nearest neighbor. The difference between each attribute of the selected instance and that of its nearest neighbor is calculated and then multiplied by a random number ranging from 0 to 1. Finally, a new instance is generated by adding the difference to the respective attribute values of the selected instance. For illustration, let's assume an instance i having attributes (i_{a1}, i_{a2}) = (6, 4), its nearest neighbor n having attributes (n_{a1} ,n_{a2}) = (4, 3), and a random number of 0.1. The difference between two instances can be calculated and multiplied by the random number, thus resulting in (d_{a1} ,d_{a2}) = (-2*0.1, -1*0.1) = (-0.2, -0.1). That is, a new instance i' is generate as (i'_{a1}, i'_{a2}) = (6, 4) + (-0.2, -0.1) = (5.8, 3.9).

On the other hand, under-sampling retains the integrity of the minority by reducing the number of instances of the majority class in a training sample, hereby creating a balanced sample across different outcome classes. Under-sampling randomly samples the instances to be removed and thus may incur the risk of removing the highly representative instances from the training sample. To address this problem, Kubat and Matwin (1997) propose one-sided selection algorithm, which target the redundant, boundary, and noisy instances for removals rather than relying on random sampling totally. To alleviate the negative effects of sampling bias, previous research also employs a multi-classifier committee through repeated sampling (Chan et al. 1999). In general, a multi-classifier committee uses a set of classifiers to jointly predict a new instance. To induce each classifier, the number of instances to be randomly selected from the majority class equals that of all the minority-class instances in the training sample. An inductive learning technique is then applied to these sampled instances (of the majority class) as well as all the instances of the minority class, hereby constructing an automated classifier. This process is repeated for a specified number of times to create a classifier committee.

Although resampling can address the imbalanced training sample problem through random sampling, it may not be appropriate to construct a classifier for acute appendicitis because of the diverse and atypical symptoms. The duplicated or synthesized instances created by over-sampling may not sufficiently represent the cases. On the other hand, the effectiveness of under-sampling is constrained by the quality of random sampling. Sampling evenly from the population (i.e., the majority class) may make the training sample more representative, and thus, enhancing the classifier's prediction accuracy and generalization. On the contrary, poor sampling may introduce prediction biases in the classifier. In this study, we address the imbalanced training sample problem by proposing a pre-clustering-based approach that clusters the instances of the majority class into a set of groups before randomly sampling the training examples from the majority-class instances. Our approach may create a more representative training sample by evenly sampling from the majority-class instances and hereby improves and stabilizes the prediction accuracy of the classifier constructed. For comparison purposes, we include in our evaluation traditional SVM, an over-sampling technique (i.e., SMOTE), and a multi-classifier committee built on under-sampling as benchmark techniques.

3. PRE-CLUSTERING-BASED CLASSIFICATION TECHNIQUE

We take a pre-clustering-based approach to address the skewed training sample problem in acute appendicitis diagnosis. Specifically, we develop a pre-clustering-based classification (PCC) technique, which adopts a under-sampling strategy by reducing the number of instances pertaining to the majority class (i.e., positive acute appendicitis) in the training sample. By pre-clustering the similar examples of the majority class into groups and then selecting representative examples accordingly, our proposed technique can reduce the information loss resulted from random sampling. In the following, we detail the design of the PCC technique

As shown in Figure 1, the overall processing of the PCC technique consists of three phases: majority example pre-clustering, representative example selection, and inductive classifier learning. In the majority example pre-clustering phase, the proposed technique clusters the training examples of the majority class (i.e., positive appendicitis) into a specified number of groups in order to simplify the algorithm and avoid tedious parameter tuning. For lower variances within each group and reduced the potential biases resulted from random sampling, the PCC favors a small-sized group in which the training examples are more similar to one anther, compared with those of a large-sized group. By making the number of groups to be clustered equal to the number of examples of the minority class (i.e., negative appendicitis), we can balance the instances pertaining to each outcome class while creating a large number of condensed clusters (groups). To partition the instances of the majority class, the PCC employees the *k*-means clustering algorithm (Huang 1998), which commences by randomly selecting *k* instances as the centroid of *k* distinct groups and swapping instances among groups until reaching a minimal total squared error.

In the subsequent representative example selection phase, the PCC randomly draws an example from each of the clusters created from the prior phase as its representative and use the selected examples as representative training examples of the majority class. Finally, an inductive classifier learning algorithm is applied to these representative examples of the majority class, in conjunction with the training examples of the minority class, to construct a learning classifier for predicting acute appendicitis. Specifically, the PCC employs SVM, a popular linear classifier learning algorithm for prediction model building.

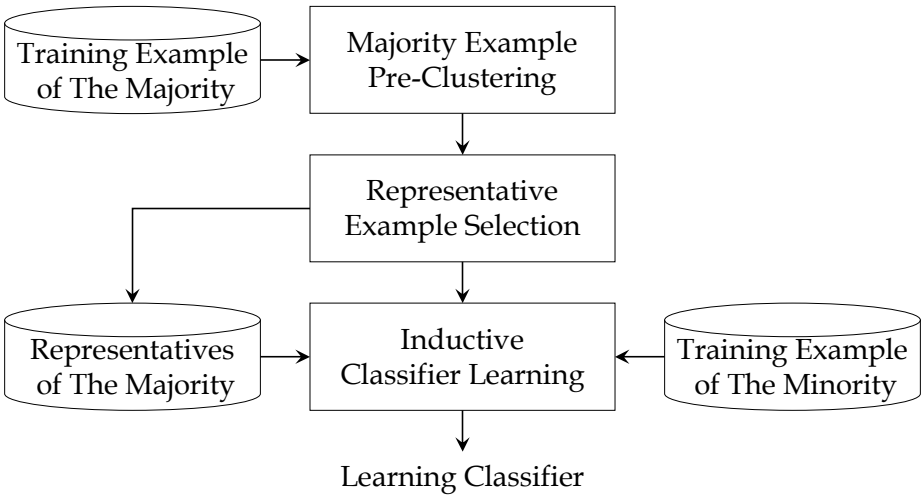

Figure 1. Overall Processing of Pre-Clustering-based Classification (PCC) Technique

4. EMPIRICAL EVALUATION AND RESULTS

We empirically evaluated the effectiveness of the PCC technique, using clinical cases obtained from a comprehensive medical center in Taiwan. Our evaluation also included Alvarado scoring system, inductive learning algorithm (i.e., SVM), over-sampling approach (i.e., SMOTE), a multiple-classifier committee for performance benchmark purposes. In the following, we detail our data collection, evaluation design, and important evaluation results.

4.1 Data collection

We collected a total of 716 patient examination reports of appendectomy from a tertiary medical center located in southern Taiwan, between January 2003 and July 2005. Each case consisted of several attributes that included the patient's age, gender, temperature, C-reactive protein (CRP), white blood cell count (WBC), segment form (SEG), migration of abdominal pain, anorexia, nausea or vomiting, right lower quadrant pain (RLQ Pain), rebound tenderness. To obviate the potential influences of other factors (e.g., other illnesses or diseases), we first removed the cases that had a documented delayed appendectomy as well as those with multiple surgical procedures performed simultaneously; e.g., acute appendicitis and other medical conditions. We also removed incomplete cases, including those with missing values. As a result, our sample consisted of 574 clinical cases; among them, 110 cases pertained to negative appendectomy confirmed by pathological examinations of the dissected specimen. In Table 1, we provide some descriptive statistics of our sample.

(a) Descriptive Statistics of Nominal Attributes

Attribute	YES	NO
Migration of Abdominal Pain	194	380
Anorexia	190	384
Nausea or Vomiting	201	373
RLQ Pain	555	19
Rebound Pain	387	187
Evaluation of Temperature	144	430
Sex (M/F)	323	251
Acute Appendicitis	464	110

(b) Descriptive Statistics of Numerical Attributes

Attribute	Average	Minimum	Maximum	Positive Appendicitis	Negative Appendicitis
Temperature (°C)	36.95	35.1	40.5	36.95	36.97
CRP (mg/L)	48.38	0.2	243.1	48.26	48.9
WBC (No./μL)	13372.99	1300	30600	13564	12565
SEG (%)	79.6	15	99	80.56	75.57
Age	36.18	3	87	36.97	32.5

Table 1. Descriptive Statistics of the Sample Used in Our Study

4.2 Evaluation Design and Measurements

To mitigate the influences of prediction biases on prediction accuracy, we created a testing dataset that had a balanced distribution between the number of instances pertaining to the positive and negative classes by randomly selecting 20% of the negative instances and an identical number of positive instances. The remaining instances constituted a training dataset that had a skewed distribution between positive and negative appendectomy, with an approximate 5-to-1 ratio in favor of positive appendectomy. The training dataset is appropriate for our evaluation purposes, since the percentage of positive cases is sufficient to make the prediction of traditional inductive learning algorithms favor to that class. To minimize the potential biases resulting from random sampling, we expanded the number of trails by repeating this process 30 times. We hereafter report the effectiveness of each investigated technique on the basis of its average performance across the 30 random trials.

To evaluate the effectiveness of each technique with respect to an outcome class (positive or negative acute appendicitis), we employed sensitivity and specificity measurements that are commonly used in clinical diagnoses. Sensitivity measures the percentage of actual positive cases that are correctly identified as positive, whereas specificity measures the proportion of actual negative cases that are correctly identified as negative. In addition, we also used *g-criterion*, the geometric mean of sensitivity and specificity, which provides a holistic measurement by considering both sensitivity and specificity rates while taking into account their trade-off. This geometric mean is less affected by extreme values than the arithmetic mean and is useful in the presence of positively skewed distributions. In general, a higher *g-criterion* suggests a classification technique achieves a higher predictive accuracy while maintaining a greater balance between the sensitivity and specificity rates (Kubat & Matwin 1997). Formally, the measurements included in our evaluation are defined as follows: $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, and $g\text{-}criterion = \sqrt{Sensitivity \times Sepcificity}$, where *TP* represents the number of testing examples correctly classified as of the positive class, *TN* denotes the number of testing examples correctly classified as of the negative class, *FP* is the number of testing examples incorrectly classified as of the positive class, and *FN* indicates the number of testing examples incorrectly classified as of the negative class.

4.3 Evaluation Results

In our evaluation, we compared the effectiveness of the proposed PCC technique and several prevalent benchmark techniques that include Alvarado scoring system, traditional SVM algorithm,

SMOTE (a common over-sampling technique), and a multi-classifier committee. To evaluate the Alvarado scoring system, we followed the suggestion by Alvarado (1986) by adopting seven as the minimum threshold score for acute appendicitis. That is, we considered cases with a total score of seven or higher acute appendicitis, and non-acute appendicitis otherwise. To assess the SMOTE, we expanded the number of the minority examples by four times to balance against that of the majority class. Moreover, we adopted a under-sampling technique to create a balanced training dataset for constructing a multi-classifier committee. We performed parameter-tuning experimentally to determine an appropriate number of classifiers, thereby forming a committee of nine classifiers to predict acute appendicitis. All the investigated techniques that follow a sampling approach employed the SVM algorithm as a basis for building their respective classifiers.

As we summarize in Table 2, there exists a trade-off between the sensitivity and specificity rates associated with each investigated technique. The SVM showed a significant prediction bias in favor of the positive class, thus resulting in the highest sensitivity rate of 1 and the lowest specificity rate of 0. The SMOTE, an over-sampling technique, also exhibited a tendency of favoring the positive class, as suggested by a high sensitivity rate of 0.68 and a relatively low specificity rate of 0.36. The prediction biases seemed to reduce the overall predictive accuracy of the SVM as well as the SMOTE, which respectively had the lowest *g-criterion* scores of 0 and 0.48 among all the investigated techniques. In contrast, the sensitivity and specificity rates achieved by the Alvarado scoring system, the multi-classifier committee, and the proposed PCC technique appeared more balanced as they attained a *g-criterion* score higher than that achieved by the SVM or the SMOTE.

Furthermore, the PCC technique achieved the highest *g-criterion* score of 0.58, seemingly arriving at a better overall prediction accuracy than any of the techniques we investigated. Specifically, the PCC technique achieved a sensitivity rate of 0.59 and a specificity rate of 0.57, both of which were higher than those associated with the multi-classifier committee. Because both the PCC technique and the multi-classifier committee employ under-sampling, our comparative results suggested the proposed pre-clustering-based technique advantageous for drawing representative instances from the majority-class instances. Moreover, in comparison with the Alvarado scoring system, the PCC technique seemed more sensitive in identifying positive acute appendicitis while being able to maintain a comparable specificity with respect to the negative cases. Overall, our proposed technique appeared to address the skewed distribution problem effectively, capable of reaching higher sensitivity and specificity rates, as compared with the benchmark techniques we studied.

	Sensitivity	**Specificity**	**g-criterion**
PCC	0.59	0.57	0.58
Alvarado Scoring System	0.50	0.60	0.54
SVM	1.00	0.00	0.00
SMOTE	0.68	0.36	0.48
Multi-classifier Committee	0.58	0.56	0.56

Table 2. Comparative Evaluation Results

5 CONCLUSION

Acute appendicitis is a common medical condition that requires a timely, accurate diagnosis that is often challenged by the diverse and atypical symptoms associated. Prior research shows the value of classification analysis for acute appendicitis diagnosis, toward which the skewed distribution of the instances pertaining to positive and negative acute appendicitis has been problematic. In this study, we propose a pre-clustering-based classification (PCC) technique to address the skewed distribution problem. According to our comparative evaluation results, the PCC seems to outperform the benchmark salient techniques, as manifested by its higher sensitivity, specificity, and *g-criterion* scores. That is, our proposed pre-clustering-based technique is capable of making accurate predictions of acute appendicitis without favoring the positive or negative class.

This study has contributed to the imbalanced learning sample problem in general and acute

appendicitis diagnosis support in particular. The proposed PCC technique is able to alleviate the negative influences of an imbalanced sample on prediction accuracy. We empirically examine the effectiveness of the classifier constructed by the proposed technique, using prevalent practice and classification techniques built on resampling as performance benchmarks. Moreover, the PCC technique is effective and naïve without having to tune any additional parameter further, and thus, can be easily implemented and applied. Overall, our evaluation results show that the PCC technique is less biased toward the positive or negative class, compared with the benchmark techniques. The imbalanced sample problem exists in many real-world situations, above and beyond healthcare management. For example, the proposed PCC technique can be extended to cope with the imbalanced sample problem associated with fraud detection, intrusion detection, product recommendation, spam mail filtering, and etc..

Our study has several limitations that in turn point to several important future research directions. First, we assume the availability of the attribute values in each training example, which may not hold in some prediction scenarios often characterized by incomplete records or missing attribute values. It is essential to cope with such problems by extending the PCC technique. Second, the dataset used in current study is limited in terms of sheer volume and source as they are obtained from a single medical center. Additional data collections, preferably from other healthcare organizations, are necessary for further evaluating the effectiveness of the proposed technique, hereby generating more convincing evidence suggesting its clinical efficacy. Third, we show the use of the PCC technique to support acute appendicitis diagnosis; for increased generalizability, its applications in classification scenarios involving similar skewed distribution problems also have to be examined.

References

Alvarado, A. (1986). A Practical Score for the early Diagnosis of Acute Appendicitis. Annals of Emergency Medicine, 15, 557-564.

Abdeldaim, Y., Mahmood, S. and Mc Avinchey, D. (2007). The Alvarado Score as a Tool for Diagnosis of Acute Appendicitis. Irish Medical Journal, 101 (1), 342.

Bijnen, C.L., van den Broek, W.T., Bijnen, A.B., de Ruiter, P. and Gouma, D.J. (2003). Implications of Removing a Normal Appendix. Digestive Surgery, 20, 215-221.

Birnbaum, B.A. and Wilson, S.R. (2000). Appendicitis at the Millennium. Radiology, 215, 337-348.

Clark, P. and Boswell, R. (1991). Rule Induction with CN2: Some Recent Improvements. In Proceedings of the fifth European Working Session on Learning, p. 151-163, Porto, Portugal.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (JAIR), 16, 321-357.

Chan, P.K., Fan, W., Prodromidis, A.L., and Stolfo, S.J. (1999). Distributed Data Mining in Credit Card Fraud Detection. IEEE Intelligent Systems, 14 (6), 67-74.

Cheng, T.H. and Hu, P. (2009). A Data-Driven Approach to Manage the Length of Stay for Appendectomy Patients. IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 39, 1339-1347.

Dasarathy, B. V. (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. McGraw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, CA.

DeRouin, E., Brown, J., Beck, H., Fausett, L., and Schneider, M. (1991). Neural Network Training on Unequally Represented Classes. Intelligent Engineering Systems Through Artificial Neural Networks, ASME Press, New York, 135-145.

Ditillo, M.F., Dziura, J.D. and Rabinovici, R. (2006). Is It Safe to Delay Appendectomy in Adults With Acute Appendicitis?. Annals of Surgery, 244, 656-660.

Delgado, M., Sánchez, D., Martín-Bautista, M.J., and Vila, M. (2001). Mining Association Rules with Improved Semantics in Medical Databases. Artificial Intelligence in Medicine, 21, 241-245.

Filewood, F. (2006). Improving diagnosis and treatment for appendicitis. Nursing times, 101, 41.

Flum, D.R. and Koepsell, T. (2002). The Clinical and Economic Correlates of Misdiagnosed Appendicitis: Nationwide Analysis. Archives of Surgery, 137, 799-804.

Graffeo, C.S. and Counselman, F.L. (1996). Appendicitis. Emergency Medicine Clinics of North America, 14, 653-671.

Heckerman, D. (1997). Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery, 1 (1), 79-119.
Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 2, 283-304.
Hardin, D.M. (1999). Acute Appendicitis: Review and Update. American Family Physician, 60, 2027-2034.
Kraft, M.R., Desouza, K.C. and Androwich, I. (2003). Data Mining in Healthcare Information Systems: Case Study of a Veteran's Administration Spinal Cord Injury Population. In Proceedings of the 36th Hawaii International Conference on System Sciences, p. 159.
Kubat, M., Holte, R., and Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning, 30, 195–215.
Al-Khayal, K.A. and Al-Omran, M.A. (2007). Computed Tomography and Ultrasonography in the Diagnosis of Equivocal Acute Appendicitis: A meta-analysis. Saudi Med J, 28 (2), 173–180.
Kubat, M., and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, p. 179–186.
Liang, M.K. (2005). The Art and Science of Diagnosing Acute Appendicitis. Southern Medical Journal, 98 (12), 1159-1160.
Lewis, D., and Catlett, J. (1994). Heterogeneous Uncertainty Sampling for Supervised Learning. In Proceedings of the Eleventh International Conference of Machine Learning (Cohen, W. & Hirsh, H Eds.), p.148–156, San Francisco, CA.
Mun, S., Ernst, R.D., Chen, K., Oto, A., Shah, S. and Mileski, W.J. (2006). Rapid CT Diagnosis of Acute Appendicitis with IV Contrast Material. Emergency Radiology, 12, 99-102.
Japkowicz, N. (2000). Learning from Imbalanced Data Sets: A Comparison of Various Strategies. In Proceedings of the AAAI Workshop on Learning from Imbalanced Data Sets, p. 10-15.
Ng, S.P., Cheng, S.M., Yang, F.S., Tzen, C.Y., and Huang, J.K., (2007). Hyperdense Appendix on Unenhanced CT: A Sign of Acute Appendicitis. Abdominal Imaging, 32 (6), 701–704.
Old, J.L., Dusing, R.W., Yap, W. and Dirks, J. (2005). Imaging for Suspected Appendicitis. American Family Physician, 71 (1), 71-78.
Prabhudesai, S.G., Gould, S., Rekhraj, S., Tekkis, P.P., Glazer, G., and Ziprin, P. (2008). Artificial Neural Networks: Useful Aid in Diagnosing Acute Appendicitis. World Journal of Surgery, 32, 305-309.
Pieper, R., Kager, L. and Nasman P. (1982). Acute Appendicitis: A Clinical Study of 1018 Cases of Emergency Appendectomy. Acta Chirurgica Scandinavica, 148, 51-62.
Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning, 1, 81-106.
Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.
Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning Internal Representations by Error Propagation. Parallel Distriduted Processing: Explorations in the Microstructures of Cognition (Rumelhart, D.E. & McClelland J.L. Eds.), MIT Press, Cambridge, MA, 1, 318-362.
Schwartz, S.I. (1994). Appendix, Principles of Surgery. 6th Edition. McGraw Hill, New York, 1994.
Terasawa, T., Blackmore, C.C., Bent, S., and Kohlwes R.J. (2004). Systematic Review: Computed Tomography and Ultrasonography to Detect Acute Appendicitis in Adults and Adolescents. Ann Intern Med, 141, 537-546.
Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York, 1995.
Walczak, S. and Scharf, J.E. (2000). Reducing Surgical Patient Costs Through Use of An Artificial Neural Network to Predict Transfusion Requirements. Decision Support Systems, 30 (2), 125-138.
Zorman, M., Eich, H.P., Kokol, P., and Ohmann, C. (2001). Comparison of Three Databases with A Decision Tree Approach in the Medical Field of Acute Appendicitis. Health Technology and Informatics, 84 (2), 1414-1418.