Asian Journal of Computer Science and Technology
ISSN 2249-0701 Vol. 1 No. 1, 2012, pp.1-5

Data Pre-Processing for Classification and Clustering

S.Balamurugan and A.B.Arockia Christopher
Department of Information Technology, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India
sbit@tce.edu, abac@tce.edu

***Abstract* — In real world datasets, lots of redundant and conflicting data exists. The performance of a classification algorithm in data mining is greatly affected by noisy information (i.e. redundant and conflicting data). These parameters not only increase the cost of mining process, but also degrade the detection performance of the classifiers. They have to be removed to increase the efficiency and accuracy of the classifiers. This process is called as the tuning of the dataset. The redundancy check will be performed on the original dataset and the resultant is to be preserved. This resultant dataset is to be then checked for conflicting data and if they will be corrected and updated to the original dataset. This updated dataset is to be then classified using a variety of classifiers like Multilayer perceptron, SVM, Decision stump, Kstar, LWL, Rep tree, Decision table, ID3, J48 and Naïve Bayes. The performance of the updated datasets on these classifiers is to be found. The results will show a significant improvement in the classification accuracy when redundancy and conflicts are to be removed. The conflicts after correction are to be updated to the original dataset, and when the performance of the classifier is to be evaluated, great improvement is to be witnessed.**

***Keywords* — data mining, classification algorithm, redundancy, conflicting data.**

I. Introduction

Data mining is the process of finding interesting patterns in data. Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. It often involves datasets with a large number of attributes. Many attributes in most real world data are redundant and/or simply irrelevant to the purposes of discovering interesting patterns. The performance of a classification algorithm in data mining is greatly affected by the noisy information (i.e.redundant and conflicting data). These parameters not only increase the cost of mining process, but also degrade the detection performance of the classifiers. This process is called as the tuning of the dataset.

Data mining will be done to remove redundant data and to correct the conflicting data. This is to be done in two phases. The data tuning will mainly done to improve the quality of the dataset. This consequently improves the reliability on the dataset. Since the quality of the dataset has been improved, the accuracy and performance of the classifiers that are applied on these datasets also improve.

The phase one of the data tuning process will remove the redundancy and inconsistency in the collected dataset. Removal of redundancy is important because it not only increases the cost but also degrades the performance. The conflicting datasets have same attribute values for the predicted attributes but have different attribute values for the class attributes. Classifier generally learns improperly form these conflicting data and hence the performance of the classifier degrades.

The phase two of the data tuning will be done after the phase on is completed. Classifiers mis-learn from these conflicting data and hence their performance degrades. Hence the phase two of the data tuning will correct the conflicting data and updates it to the phase one dataset so that the classifier can learn and perform well. The phase two of data tuning will correct conflicting data in dataset and it will update in dataset and it will update its to the original dataset.

II. Data Cleaning: Overview

A Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

A data cleaning approach should satisfy several requirements. First of all, it should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources. Furthermore, data cleaning should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata. Mapping functions for data cleaning and other data transformations should be specified in a declarative way and be reusable for other data sources as well as for query processing.

III. Related Work

Data preprocessing or preparation is an important and critical step in the data mining process and it has a huge impact on the success of a data mining project [20]. This goal generates

an urgent need for data analysis aimed at cleaning the raw data [56]. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right [29] & [8] propose a bit-based feature selection method to find the smallest feature set to represent the indexes of a given data set. The proposed approach originates from the bitmap indexing and rough set techniques. It consists of two phases. In the first phase, the given data set is transformed into a bitmap indexing matrix with some additional data information.

In the second phase, a set of relevant and enough features are selected and used to represent the classification indexes of the given data set. After the relevant and enough features are selected, they can be judged by the domain expertise and the final feature set of the given data set is thus proposed. Finally, the experimental results on different data sets also show the efficiency and accuracy of the proposed approach. [9] Investigate the influence of different preprocessing techniques of attribute scaling, sampling, coding of categorical as well as coding of continuous attributes on the classifier performance of decision trees, neural networks and support vector machines.

[35] Mention that data reduction is an important issue in the field of data mining. The goal of data reduction techniques is to extract a subset of data from a massive data set while maintaining the properties and characteristics of the original data in the reduced set. This allows an otherwise difficult or impossible data mining task to be carried out efficiently and effectively.

Pizzi and Pedrycz present a new methodology, which systematically addresses these design classification issues in their study. At the preprocessing phase they offer a new approach of stochastic feature selection. This type of feature selection collates quadratic ally transformed feature subsets for presentation to a collection of respective classifiers[42].

A. Noise Handling and Noise Impact Analysis

Many of these learning models involve unique noise-handling routines or structures. For example, one representative method in classification tree induction is the tree pruning, including prepruning and postpruning approaches[25]. As a representative instance-based algorithm, IB3 [12] involves a noise-tolerate strategy, in which a selective utilization filter is applied to prevent noisy instances from being selected as prototypes. In [5], the authors presented a linear complexity instance selection algorithm by performing several rounds of instance selection on subsets of the original dataset to address the scaling issue of instance selection on very large datasets. In [4], the authors proposed a fuzzy support vector machine design for learning from noisy data.

Different from this strategy, another proactive idea is to adopt data preprocessing techniques, such as noise cleansing [6,45], erroneous attribute value detection [7,53,48,10], missing attribute value acquisition [51,52], and data imputation [21], to improve the data quality before prediction models can be formed from the preprocessed data. If training and test data both suffer from the same level of noises, it has been observed that ‘‘for higher noise levels, the performance of a correct decision tree on corrupted test data was found to be inferior to that of an imperfect decision tree formed from data corrupted to a similar model’’ [26,27].

In our previous work [50], we empirically studied the impact of class noise and attribute noise to the learning modules and concluded that class noise is more harmful, than attribute noise, for prediction models. In [36], the authors empirically studied the classification algorithms for test data containing under or overly represented attribute noises and concluded that over representative (high training noise and low test noise) attribute noise has an negative impact on the learning model, while under representative (low training noise and high test noise) attribute noise is less of a concern. Conducting data cleaning on test data usually improves performance.

B. Classifier Ensembling for Data Imperfections

Ensemble learning is a proven effective tool that generally outperforms single models [47,33,1]. One of the potential advantages of classifier ensemble model is that it is shown to be robust to datasets containing noisy or missing attribute values [38]. In this subsection, we regard classifier ensembling as a treatment for data imperfections and review related work from the following two aspects: (1) building classifier ensembling from noisy data in general; (2) improving classifier ensembling methods through the enhancement of the accuracy and diversity of the base learners.

1) Classifier Ensembling for Noisy Data

Assuming the training data is noisy, whereas the test data is noise free, the prediction accuracies of general classifier ensembling methods are not always superior to the accuracies of single models, depending on how the ensemble is designed. For example, previous experiments on classification noise have shown that bagging ensemble is quite robust, whereas Boosting is more sensitive to noisy data [46]. Many evidences have shown that commonly used boosting algorithms such as AdaBoost [54] and LogitBoost [22] are sensitive to data errors and perform poorly on noisy data [24, 46, 49, 41].

In [49], the author performed a comprehensive analysis on Ada Boost in the presence of noisy data. In [41], the authors theoretically demonstrated that all the boosting algorithms that try to minimize some convex potential functions of the margins of a dataset are highly susceptible to random classification noise. A boosting algorithm called Smooth Boost has been designed to avoid assigning too much weight

to any single instance, so that the algorithm can tolerate a certain level of noise [44]. In [37], the authors have proposed a binary classifier ensemble by extending the smooth boosting approach via a branch-and-bound algorithm to construct base learners in the classifier ensemble.

Furthermore, many research efforts have been made to study the reaction of ensemble learning models on different types of noisy data. For example, in [38], the performance of three classifier ensembling methods Bagging, Boosting and DECORATE have been studied for data containing missing features, class noise, and/or attribute noise. In [40], the authors investigated the problem of learning from concept drifting data streams with noise, where samples in a data stream may be mislabeled or contain erroneous values.

2) The Base Learners' Accuracy and Diversity

Classifier ensemble is also called mixtures of experts, multiple classifier systems, consensus theory, etc. [34,31,54,39]. It is shown both theoretically and experimentally that an effective classifier ensemble should consist of base learners with high-accuracy and high-diversity in predictions [47,34]. To generate a high quality classifier ensemble, one commonly agreed principle is that the ensemble should have diverse base learners. In fact, the way of measuring ''diversity'' among base learners is not unique. For example, in [47], two classifiers are considered diverse if they make different errors on new data points; and in [39], the measure of disagreement is referred as the diversity of the ensemble. In [34], the authors carries out a study on various measures of diversity in classifier ensembles. In [17], a study on the behavior of classifier ensembling methods is performed through a simulation of generating a set of classifier outputs with specified individual accuracies of base learners and fixed pair wise agreement of the measured diversity.

Examples include (1) introducing randomness into the training data by using different sampling mechanisms, such as Bagging [31], Boosting [54] and DECORATE [39]; (2) using ensemble feature selection to introduce randomness [11,16,2,3]; (3) introducing randomness into the class attributes of the training data [32,15]; and (4) adjusting parameters of the learning models to introduce randomness [23,43]. The traditional wisdom in general agrees that the accuracy of base learners should be higher than 50% [47]. The higher the accuracy the base learners could achieve, the better the classifier ensemble may be. When learning from noisy data sources, we believe that using data cleansing to construct base learners with high accuracy is important.

IV. Proposed Work

Steps involved in data cleaning are shown with the help of block diagram in Figure 1 below:

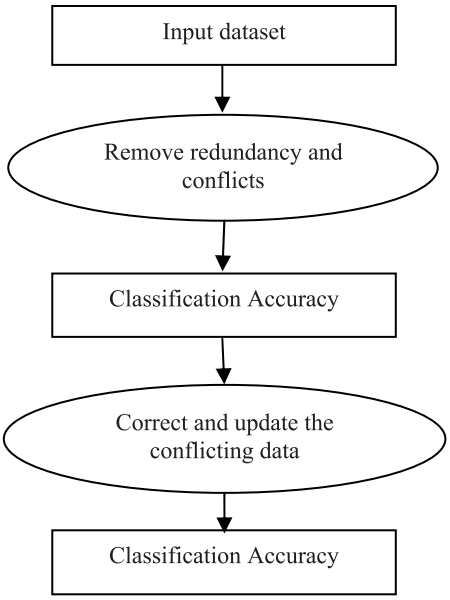

Fig.1 Block diagram for data cleaning

Procedure for the data cleaning

1. Dataset is to be provided
2. Redundancy and conflicting data in the dataset are to be removed.
3. The conflicting data sets are to be corrected and updated.
4. Accuracy of classification algorithm is to be evaluated.

In this work, data mining will be implemented on a dataset in order to increase the classification accuracy and to reduce the computational time. In general, records are said to redundant if they have same predictive attribute values and same class label values. Many of the attributes in most real world data are redundant and/or simply irrelevant to the purposes of discovering interesting patterns. Records are said to be conflicting if they have same predictive attribute values but different class label values. The classifier does not learn properly from conflicting records as they provide two class labels for the same predictive attribute values. Data mining will be done in two phases namely

Phase 1: Redundant and conflicting data will be removed from the original dataset.

Phase 2: Conflicting data will be passed as test data to the classification algorithms and its class label value is to be found. The class label value that will found is to be updated to the original training data.

The phase one of data tuning will remove redundancy and inconsistency in the collected dataset. The performance of the classifier is to be found at the end of the phase one. Phase one of the data tuning process is to be followed by Phase two. Classifiers cannot properly learn from these conflicting data and hence the performance of the classifiers degrades.

Hence, in phase two, the conflicting data will be corrected and updated to the resultant dataset of Phase one to improve the performance and accuracy of the classifier. The performance of the classifier is to be again found at the end of the phase two. A significant improvement performance at the end of the phase two is to be observed.

V. Conclusion

The cleaning step is necessary to resolve several types of problems include noisy data, redundancy data, missing data values, etc. All the classification algorithms rely heavily on the product of this stage, which is the final training set. By selecting relevant instances; experts can usually remove irrelevant ones as well as noise and/or redundant data. The high quality data will lead to high quality results and reduced costs for data mining. In addition, when a data set is too huge, it may not be possible to run a classification algorithm. In most cases, missing data should be pre-processed so as to allow the whole data set to be processed by a classification algorithm. It would be nice if a single sequence of data cleaning algorithms had the best performance for each data set but this is not happened. Thus, we presented the most well known algorithms for each step of data cleaning so that one achieves the best performance for their data set.

Acknowledgment

The authors wish to acknowledge the financial support from the University Grant Commission (UGC), INDIA for the Major Research Project “Data Tuner for effective Data Pre-processing” vide reference F.No. 39-899/2010 (SR), and also gratefully acknowledge the unanimous reviewers for their kind suggestions and comments for improving this paper.

References

[1] A. Gal, T. Sagi, Tuning the ensemble selection process of schema matchers, *Information Systems* 35 (8) (2010) 845–859.

[2] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection, *Information Fusion* 6 (1) (2005) 83–98.

[3] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Sequential genetic search for ensemble feature selection, in: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (IJCAI-2005), 2005, pp. 877–882.

[4] C. fu Lin, S. de Wang, Training algorithms for fuzzy support vector machines with noisy data, *Pattern Recognition Letters* 25 (14) (2004) 1647–1656.

[5] C. Garcı´a-Osorio, A. de Haro-Garcı´a, N. Garcı´a-Pedrajas, Democratic instance selection: a linear complexity instance selection algorithm based on classifier ensemble concepts, *Artificial Intelligence* 174 (5–6) (2010) 410–41.

[6] C.E. Brodley, M.A. Friedl, Identifying mislabeled training data, *Journal of Artificial Intelligence Research* 11 (1999) 131–167.

[7] C.M. Teng, Correcting noisy data, in: *Proceedings of the Sixteenth International Conference on Machine Learning* (ICML-1999), 1999, pp. 239–248.

[8] Chen, W., Tseng, S., & Hong, T. (2008). An efficient bit-based feature selection method. *Expert Systems with Applications*, 34, 2858–2869.

[9] Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, 173, 781–800.

[10] D. He, X. Zhu, X. Wu, Error detection and uncertainty modeling for imprecise data, in: ICTAI ’09 *Proceedings of the 2009 21st IEEE International Conference on Tools with Artificial Intelligence,* 2009, pp. 792–795.

[11] D. Opitz, Feature selection for ensembles, in: *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-1999), 1999, pp. 379–384.

[12] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithm, *Machine Learning* 6 (1) (1991) 37–66.

[13] Feyza and Lale: Data mining and preprocessing application on component reports of an airline company in Turkey. *Expert Systems with Applications* 38 (2011) 6618-6626.

[14] Friedman, J.H. 1997. Data mining and statistics: What’s the connection? *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics.*

[15] G. Martinez-Munoz, A. Suarez, Switching class labels to generate classification ensembles, *Pattern Recognition* 38 (10) (2005) 1483–1494.

[16] G. Zenobi, P. Cunningham, Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error, in: *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001),* Freiburg, Germany, 2001, pp. 576–587.

[17] H. Zouari, L. Heutte, Y. Lecourtier, Controlling the diversity in classifier ensembles through a measure of agreement, *Pattern Recognition* 38 (11) (2005) 2195–2199.

[18] Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998.

[19] http://www.google.com

[20] Hu, X. (2003). DB-reduction: A data preprocessing algorithm for data mining applications. *Applied Mathematics Letters,* 16, 889–895.

[21] I. Fellegi, D. Holt, A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association* 71 (1976) 17–35.

[22] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, *The Annals of Statistics* 28 (2) (2000) 337–374.

[23] J.F. Kolen, J.B. Pollack, Back propagation is sensitive to initial conditions, Advances in Neural Information Processing Systems 3 (1991) 860–867.

[24] J.R. Quinlan, Bagging, boosting, and c4.5, in: *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (AAAI-1996), 1996, pp. 725–730.

[25] J.R. Quinlan, Induction of decision trees, *Machine Learning* 1 (1) (1986) 81–106.

[26] J.R. Quinlan, Learning from noisy data, in: *Proceedings of the Second International Machine Learning Workshop*, 1983.

[27] J.R. Quinlan, The effect of noise on concept learning, in: R.S. Michalski, J.G. Carboneel, T.M. Mitcheel (Eds.), Machine Learning, 1986.

[28] K. M. Ho, and P. D. Scott. Reducing Decision Tree Fragmentation Through Attribute Value Grouping: A Comparative Study, in Intelligent Data Analysis Journal, 4(1), pp.1-20, 2000.

[29] Kim, Y., Street, W. N., & Menczer, F. (2003). Feature selection. In Data mining. USA:University Of Iowa.

[30] Kubat, M. and Matwin, S., ‘Addressing the Curse of Imbalanced Data Sets: One Sided Sampling’, in the *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179-186, 1997.

[31] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[32] L. Breiman, Randomizing outputs to increase prediction accuracy, Machine Learning 40 (3) (2000) 229–242.

[33] L. Rokach, Ensemble-based classifiers, *Artificial Intelligence Review* 33 (2010) 1–39.

[34] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning* 51 (2) (2003) 181–207.

[35] Li, X., & Jacob, V. S. (2008). Adaptive data reduction for large-scale transaction data. *European Journal of Operational Research,* 188, 910–924.

[36] M. Mannino, Y. Yang, Y. Ryu, Classification algorithm sensitivity to training data with non representative attribute noise, *Decision Support Systems* 46 (3) (2009) 743–751.

[37] N. Goldberg, C. Chieh Shan, N. Goldberg, C. Chieh Shan, Boosting optimal logical patterns using noisy data, in: *Proceedings of the Seventh SIAM International Conference on Data Mining,* Minneapolis, Minnesota, 2007.

[38] P. Melville, N. Shah, L. Mihalkova, R.J. Mooney, Experiments on ensembles with missing and noisy data, in: *Proceedings of the Workshop on Multi Classifier Systems,* Springer Verlag, 2004, pp. 293–302.

[39] P. Melville, R.J. Mooney, Constructing diverse classifier ensembles using artificial training examples, in: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (IJCAI-2003), 2003, pp. 505–510.

[40] P. Zhang, X. Zhu, Y. Shi, L. Guo, X. Wu, Robust ensemble learning for mining noisy data streams, *Decision Support Systems* 50 (2011) 469–479.

[41] P.M. Long, R.A. Servedio, Random classification noise defeats all convex potential boosters, in: *Proceedings of the 25th International Conference on Machine Learning* (ICML), Helsinki, Finland, 2008, pp. 608–615.

[42] Pizzi, N. J., & Pedrycz, W. (2008). Effective classification using feature selection and fuzzy. Integration Fuzzy Sets and Systems.

[43] R. Maclin, J.W. Shavlik, Combining the predictions of multiple classifiers: using competitive learning to initialize neural networks, in: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (IJCAI-1995), Montreal, Canada, 1995, pp. 524–530.

[44] R.A. Servedio, Smooth boosting and learning with malicious noise, *Journal of Machine Learning Research* 4 (2003) 633–648.

[45] S. Shah, A. Kusiak, Relabeling algorithm for retrieval of noisy instances and improving prediction quality, Computers in Biology and Medicine 40 (2010) 288–299.

[46] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139–157.

[47] T.G. Dietterich, Ensemble methods in machine learning, *Proceedings of the First International Workshop on Multiple Classifier Systems,* Lecture Notes in Computer Science, vol. 1857, 2000, pp. 1–15.

[48] T.M. Khoshgoftaar, J.V. Hulse, Empirical case studies in attribute noise detection, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*—Special Issue on Information Reuse and Integration 39 (4) (2009) 379–388.

[49] W. Jiang, Boosting with noisy data: some views from statistical theory, Neural Computation 16 (4) (2004) 789–810.

[50] X. Zhu, X. Wu, Class noise vs attribute noise: a quantitative study of their impacts, *Artificial Intelligence Review* 22 (3–4) (2004) 177–210.

[51] X. Zhu, X. Wu, Cost-constrained data acquisition for intelligent data preparation, *IEEE Transactions on Knowledge and Data Engineering* (TKDE) 17 (11) (2005) 1542–1556.

[52] X. Zhu, X. Wu, Data acquisition with active impact-sensitive instance selection, in: *Proceedings of the Sixteenth IEEE International Conference on Tools with Artificial Intelligence* (ICTAI-2004), 2004, pp. 721–726.

[53] X. Zhu, X. Wu, Y. Yang, Error detection and impact-sensitive instance ranking in noisy datasets, in: *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (AAAI-2004), 2004, pp. 378–384.

[54] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: *Proceedings of the Thirteenth International Conference on Machine Learning* (ICML-1996), 1996, pp. 148–156.

[55] Yan Zhang, Xingquan Zhu, Sindong Wu, Jeffrey P.Bond: Corrective classification: Learning from data imperfections with aggressive and diverse classifier ensembling . *Information Systems* 36 (2011) 1135-1157.

[56] Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence,* 173, 75–381.