Journal of Computer Science 5 (2): 131-135, 2009
ISSN 1549-3636

Improving Accuracy and Coverage of Data Mining Systems that are Built from Noisy Datasets: A New Model

Luai A. Al Shalabi
Faculty of Computer Studies, Arab Open University, Kuwait Branch, Kuwait

Abstract: Problem statement: Noise within datasets has to be dealt with under most circumstances. This noise includes misclassified data or information as well as missing data or information. Simple human error is considered as misclassification. These errors will decrease the accuracy of the data mining system so it will not be likely to be used. The objective was to propose an effective algorithm to deal with noise which is represented by missing data in datasets. **Approach:** A model for improving the accuracy and coverage of data mining systems was proposed and the algorithm of this model was constructed. The algorithm was dealing with missing values in datasets. It splits the original dataset into two new datasets; one contains tuples that have no missing values and the other one contains tuples that have missing values. The proposed algorithm was applied to each of the two new datasets. It finds the reduct of each of them and then it merges the new reducts into one new dataset which will be ready for training. **Results:** The results showed interesting as it increases the accuracy and coverage of the tested dataset compared to the traditional models. **Conclusion:** The proposed algorithm performs effectively and generates better results than the previous ones.

Key words: Data mining, noise, missing values, rule generation, knowledge discovery

INTRODUCTION

Data mining is a relatively new field emerging in many disciplines. It is becoming more popular as technology advances and the need for efficient data analysis is required. The aim of data mining is not to provide strict rules by analyzing the full data set, but it is used to predict with some certainty while only analyzing a small specific representative part of the data. Therefore, 'rules generated by data mining are empirical'-'they are not physical laws'[5]. Many methods of data mining exist. Some of these methods include a rule induction and a K-nearest neighbor.

Data mining is a form of machine discovery where the discovered knowledge is represented in a high level language. It is capable of discovering domain knowledge from given examples. The type of rule or pattern that exists in data depends on the domain. Discovery systems have been applied to real databases in medicine[2,6], astronomy[7], the stock market[4] and many other areas.

One common problem or challenge in data mining and knowledge discovery research is a noisy data[3,10]. In large databases, many of the attribute values are unknown because of the unavailability of data. Also, attribute values could be incorrect due to an erroneous instrument measuring some property or human error when registering it. Noisy data will definitely minimize the accuracy of any data mining system.

Missing values: Al Shalabi[10] summarized two forms of noise in the data as described:

Corrupted values: Sometimes some of the values in the training set are altered from what they should have been. This may result in one or more tuples in the data set conflicting with the rules already established. The system may then consider these values as noise and ignore them. The problem is that one never knows if these values are correct or not and the challenge is how to handle strange or unexpected values in the best manner.

Missing attribute values: One or more of the attribute values may be missing both for examples (tuples) in the training set and for examples which are to be classified[9]. Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns[1]. If attributes are missing in any training set, the system may either ignore this object totally, try to take it into account by, for instance, finding what is the missing attribute's most probable value, or use the value "missing", "unknown" or "NULL" as a separate value for the attribute.

The problem of missing values has been investigated since many times ago[8,14]. The simple solution is to discard the data instances with some

missing values[17]. A more difficult solution is to try to determine these values[13]. Several techniques to handle missing values have been discussed in the literature[3,11,12,13,14,17].

Mitchell[15] proposed that missing values within training sets can be managed by assigning values that are seen in similar cases. For example the value found that is most common when another attribute matches that of a full record. This method requires some inference into which attribute is most relevant to the missing attribute. According to Mitchell[15], another method is to assign the average of the missing attributes that correspond with another relevant attribute as above.

Mitchell[15] presents a third method, which is the method used within C4.5. The attributes which contain missing values are given probabilities for each possible value. When the missing value is being considered, the probabilities are assigned as values of a new fractional attribute weighted by considering the aforementioned probabilities and the decision tree is created as normal[9].

MATERIALS AND METHODS

An over view will be given to two important data mining models. These methods will be tested against the dataset which is used in the experiment in this study.

Rule induction: A data mine system has to infer a model from the dataset that it may define classes such that the dataset contains one or more attributes that denote the class of a record (the predicted attributes) while the remaining attributes are the predicting attributes. Class can then be defined by condition on the attributes. When the classes are defined, the system should be able to infer the rules that govern classification, in other words the system should find the description of each class.

Production rules have been widely used to represent knowledge and they have the advantage of being easily interpreted by human experts because of their modularity which means that a single rule can be understood in isolation and does not need reference to other rules. The structure of such rules can be described as if-then rules.

K-Nearest Neighbor (KNN): Dr. Kardi Teknomo[16] gave clear information in his tutorial about KNN. Some important notes about KNN are highlighted as below.

K-nearest neighbor is a supervised learning algorithm where the result of new example query is classified based on majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new example based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, K number of examples (training point) closest to the query point is found. The classification is using majority vote among the classification of the K examples. Any ties can be broken at random. K-Nearest neighbor algorithm uses neighborhood classification as the prediction value of the new query example.

K-nearest neighbor algorithm is very simple. It works based on minimum distance from the query example to the training samples to determine the K-nearest neighbors. After gathering K-nearest neighbors, simple majority of these K-nearest neighbors are taken to be the prediction of the query example.

In order to get the best solution, a maximum value for k is selected, a user builds models on all values of k up to the maximum specified value and voting is done on the best of these models.

Some advantage of using KNN technique include: robust to noisy training data, effective if the training data is large and higher value of k provides smoothing that reduces the volume of noise in the training data.

Some disadvantage of using KNN technique include: The need to determine value of parameter K (number of nearest neighbors), the distance based learning is not clear to which type of distance to use and which attribute to use to produce the best results and the computation cost is quite high because we need to compute distance of each query instance to all training samples. Also, computing time goes up as k goes up.

Proposed work: A model was proposed to deal with missing values in datasets in order to generate better accuracy and coverage values for a data mining system. An Algorithm was constructed from the model. Six steps are sequentially executed in order to get the expected results. What we needed first is to make the original dataset (ODS) under study available. ODS was divided into two different datasets: one dataset contains all examples (tuples) that do not have missing values (DS1) and the other dataset contains all examples that have missing values within each of them (DS2). Up to this point, the working space has three datasets: ODS, DS1 and DS2.

The reduct of DS1 was calculated and the attributes that compose the reduct were considered as the main important attributes. The same attributes were considered in DS2 while the others were removed from it (this is the reduct of DS2). The reduct of DS1 was

stored in the dataset which is called RDS1 and the reduct of DS2 was stored in the dataset which is called RDS2. Both reducts were merged in order to form the new dataset that will be trained in order to get the better knowledge from it. The resulted dataset was called RtTDS. A prepossessing engine is the name that was given to the above steps.

The main process is to find the conclusion (knowledge) from the RtTDS. Two different techniques were used individually to find the conclusion. These techniques are: Rule induction and KNN. A company or an institute may use one of these techniques or they may use some other techniques based on the performance of the technique itself against the nominated dataset. Here, I referred to the main rule which says that there is no perfect technique for all datasets, but each dataset is a case study by itself. One technique is the most suitable for this dataset but it is not for the others.

Figure 1 shows the traditional model of finding conclusion (knowledge) in the area of data mining and knowledge discovery while Fig. 2 shows the proposed model of finding conclusion in the area of data mining and knowledge discovery. Later in this article, a comparison between these two models will be made. The steps of the proposed model are given as an algorithm in algorithm 1.

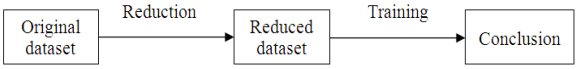

Fig. 1: The traditional model

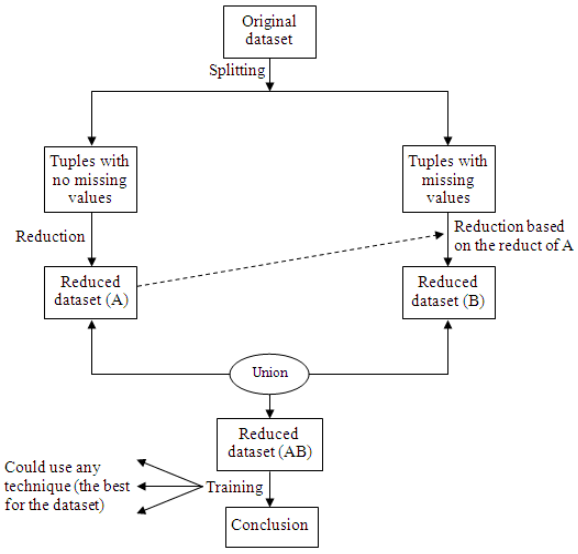

Fig. 2: The proposed model

Algorithm: The constructed algorithm:

1- Read the original dataset of n attributes (ODS).
2- Divide the original dataset into two subsets:
   a. The first subset includes only tuples without missing values of n attributes (DS1) such that:
      DS1 ← $\sigma_{(A1 \neq NULL\ AND\ A2 \neq NULL\ AND\ \ldots\ AND\ an \neq NULL)}$ (ODS)
   b. The second subset includes only tuples with missing values of n attributes (DS2).
      DS2 ← ODS-DS1
3- Find the reduct of the first dataset (DS1) such that:
   RDS1 ← RED(DS1)
4- Reduce DS2 by keeping only the attributes that were resulted from the reduct of DS1 and store the result in RDS2.
5- Merge the reduced datasets (RDS1 and RDS2) into a ready-to-train dataset (RtTDS) such that :
   RtTDS ← RDS1 U RDS2
6- Use different techniques of training in order to find the conclusion such that:
   Conclusion ← Train(RtTDS)
   //The following two techniques are only used for testing, any other technique could be used here.
   a. Find the conclusion based on Rule Induction such that:
      $(Conclusion)_{RI}$ ← $Train_{RI}$ (RtTDS)
   b. Find the conclusion based on K-Nearest Neighbor (KNN) such that:
      $(Conclusion)_{KNN}$ ← $Train_{KNN}$ (RtTDS)

RESULTS

RSES was used as a tool that conducts the accuracy and coverage of the dataset. Two different techniques within RSES were considered as stated in Table 1 and 2. For each technique, the comparison based on accuracy and coverage values was established between the Traditional Model After the Reduct (TMAR) and the Proposed Model After the Reduct (PMAR). All results are shown in Table 1 and 2.

Figure 3 and 4 show the results of comparing the traditional model to the proposed model. In Fig. 3, accuracy is taken as X-axis and the techniques are taken as Y-axis. While in Fig. 4, coverage is taken as X-axis and the techniques are taken as Y-axis.

Table 1: Results and comparisons based on accuracy

The technique	TMAR	PMAR
Using rule induction	93.1	97.9
Using KNN	83.1	95.7

Table 2: Results and comparisons based on coverage values

The technique	TMAR	PMAR
Using rule induction	98.7	100
Using KNN	100.0	100

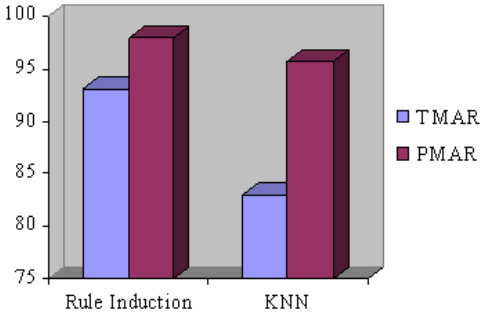

Fig. 3: Accuracy comparison of TMAR and PMAR

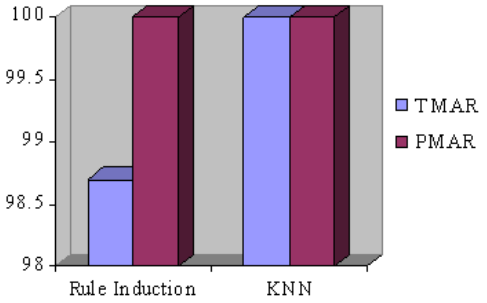

Fig. 4: Coverage comparison of TMAR and PMAR

DISCUSSION

In order to test the performance of the proposed model, a comparison was made between the results of the proposed model and the results of the traditional model.

The comparison was made based on the accuracy and the coverage resulted from each model. It was made against the Hepdata dataset which contains 468 examples and 21 attributes including the classification class (Degree = 21). 192 examples are complete (i.e., they do not have any missing values). The rest of the examples are not complete. Not complete examples have missing values that are varying between 1 and many missing values in each example.

As mentioned before in this study, Hepdata dataset was split into 2 sub datasets: DS1 and DS2. The reduct was calculated for DS1 and it was the following attributes: {1, 3, 6, 9, 10, 12, 20}. These attributes were considered as the reduct of DS2 and the other attributes were discarded. The union between the result of DS1 (all examples after the reduct) and the result of DS2 was conducted and saved into the new dataset which is called Ready-to-Train Data Set (RtTDS). RtTDS is considered of best value which is most suitable for knowledge extraction. The degree (number of attributes) of RtTDS dataset including the classification attribute is (8).

The proposed model gave accuracy value greater than the traditional model in both techniques used.

Coverage was calculated for the traditional model and the proposed model. The proposed model gave coverage values greater than that of traditional model when rule induction technique was used. KNN technique gave the same coverage value for both models.

CONCLUSION

Improving accuracy and coverage of data mining systems is a challenge. It has been noted that building data mining systems from noise data seems harder than that of cleaned data. In this article, a model was proposed and an algorithm was built that increasing the accuracy and coverage of data mining systems. The algorithm deals with existing missing values in the dataset under study. The dataset was split into two datasets based on examples that are clean (examples which do not have missing values) and those that have missing values. The reduct of the first dataset was generated by RSES. The reduct of the second dataset was formed in a specific way based on the reduct of the first dataset. The two generated reducts were merged into RtTDS and the accuracy and coverage were calculated.

Based on the previous results, the claim of this study which says that the proposed model gives better results than the traditional model is proved.

ACKNOWLEDGEMENT

This study has been supported by the Information Technology and computing program/faculty of computer studies at the Arab Open University. The author likes to thank many anonymous people for their efforts in improving the readability of this study including the patience of my wife and kids.

REFERENCES

1. Agrawal, A. and R. Srikant, 2000. Privacy preserving data mining ACM SIGMOD Rec., 29: 439-450. DOI: 10.1145/335191.335438
2. Pieter, A. and Z. Dolf, 1996. Data Mining. 1st Edn., Addison-Wesley Professional, Harlow, England, ISBN: 10: 0201403803, pp: 176.
3. Ragel, A. and B. Cremilleux, 1999. MVC: A preprocessing method to deal with missing values. Knowl. Based Syst. J., 12: 285-291. DOI: 10.1016/S0950-051(99)00022-2

4. White, A.P., 1987. Probabilistic Induction by Dynamic Path Generation in Vertual Trees. In: Research and Development in Expert Systems III, Bramer, M.A. (Ed.). Cambridge University Press, ISBN: 0-521-34145-X, pp: 35-46.
5. Read, B.J., 2000. Data mining and science? http://www.ercim.org/publication/ws-proceedings/12th-EDRG/EDRG12_Re.pdf
6. Merz, C.J. and P.M. Murphy, 1996. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html
7. Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. 1st Edn., Morgan Kaufmann, San Francisco, USA., ISBN: 10: 1558604898, pp: 500.
8. Quinlan, J.R., 1989. Unknown attribute values in induction. Proceedings of the 6th International Workshop on Machine Learning, (IWML'89), Ithaca, New York, United States, pp: 164-168. http://portal.acm.org/citation.cfm?id=102173
9. Quinlan, J.R., 1985. Induction of decision trees. Centre Adv. Comput. Sci., 1: 81-106. DOI: 10.1007/BF00116251
10. Al Shalabi, L. *et al*., 2006. Data mining: A preprocessing engine. J. Comput. Sci., 2: 735-739. http://findarticles.com/p/articles/mi_m0VVT/is_9_2/ai_n24997990
11. Al-shalabi, L., 2000. New learning models for generating classification rules based on rough set approach. Doctoral dissertation, University Putra Malaysia.
12. Al-shalabi, L. *et al*., 1999. Data mining: An overview. Proceeding of the World Engineering Congress, (WEC'99), Kuala Lumpur, Malaysia, pp: 229-234.
13. Kerdprasop, N., K. Kerdprasop, Y. Saiveaw and P. Pumrungreong, 2003. A comparative study of techniques to handle missing values in the classification task of data mining. Proceeding of the 29th Congress on Science and Technology of Thailand, Oct. 20-22, Khon Kaen University, Thailand, pp: 1-3. http://sutlib2.sut.ac.th/Sut_Article/Nittaya/BIB970_F.pdf
14. Little, R.J.A. and D.B. Rubin, 1987. Statistical Analysis with Missing Data. 1st Edn., John Wiley and Sons, ISBN: 10: 0471802549, pp: 304.
15. Mitchell, T.M., 1997. Machine Learning. 1st Edn., McGraw Hill, New York, ISBN: 10: 0071154671, pp: 352.
16. Kadri, T., 2004. Neighbors tutorial. http://people.revoledu.com/kardi/tutorial/KNN/index.html
17. Liu, W.Z., A.P. White, S.G. Thompson and M.A. Bramer, 1997. Techniques for dealing with missing values in classification. Proceeding of the 2nd International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data, Aug. 4-6, Springer-Verlag London, UK., pp: 527-536. http://portal.acm.org/citation.cfm?id=743827