International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Detection of Classifiers Using Tidbits

Margaret Flora B[1], Siva Ganesh R[2]

[1]Department of Computer Science Engineering, ST. Joseph University College of Engineering and Technology, Dar Es' Salaam, and P.OBox 11007, Tanzania

[2]Department of Electronics and Electrical Engineering, ST. Joseph University College of Engineering and Technology, Dar Es' Salaam, and P.OBox 11007, Tanzania

Abstract: ***Data mining is the computational process of discovering patterns in a collection of large data sets. It is not based on finding exact patterns alone, but it underlies the importance of tidbits. It is nothing but the collection of accurate data which is for improving the accuracy. Here the prediction of desired pattern is based on the classification algorithm. Classification is to accurately predict the target class for each case in the data. This concept can be applied for medical applications especially for cancer treatment. In this paper classification model could be used to identify the classifiers through symptoms. With respect to medical applications pattern recognition is important for the diagnosis of diseases and identification of each stages of the diseases. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. With the help of Bayes theorem, this work also identifies certain properties of tidbits, which improves the classification accuracy. This concept is used to take accurate decisions when undergoing the treatments. Experimental results can be used to predict the accuracy using the score obtained during the classification and then finds the stage of a patient undertaking cancer treatment assuming medicine as the data set for further prevention.***

Keywords: tidbits; data set; Cscore; classification accuracy;

1. Introduction

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. In data mining a lot of work has been devoted to find interesting patterns. Not much work has focused on finding exact information. Underlying the importance of finding the exact information, methodology mainly focused on finding exact information using tidbits. This work also identifies certain properties of tidbits, which improves the classification accuracy. It has several applications including detecting and identifying network intrusions. Not much work has focused on finding tidbits. Tidbits of information can take the following form during classification tasks: - small subsets of data instances that lie very close to the class boundary and are sensitive to small changes in attribute values. The magnitude of changes to the original model provides clues to the criticality of such data instances. The main objective is to find the classifiers through symptoms in identifying a stage of a patient, through current medicine. The research was extended to the mining of outliers and the concept of distance-based outliers was proposed to identify records that are different from the rest of the data set. Assuming medicine as a data set, classify the tidbits among several data. As there are number of tools available we can consider only few tools and indicate major case study in health problems. By considering the cancer data sets while 70 to 80% results can be obtained. During classification tasks the following problems arise:i)small sub elements of data instances that fall very close to the boundary and are sensitive to small changes in attribute values,those small changes result in switching of stages.ii)the accuracy is the big challenge.iii)the inability to classify accurately the data that are near the boundary.

A. Problem Definition

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

2. Identifying Tidbits

In recent times, detecting an outliers has emerged as an important area of work in the field of Data Mining. Apart from many applications in data mining the main objective is to identify tidbits. Identifying tidbits is nothing but finding the exact piece of information. It is not completely based on finding exact information rather it is to improve the accuracy by classifiers. While identifying tidbits of information, it will take the following form classification tasks: small subsets of data instances that lie very close to the class boundary and are sensitive to small changes in attribute values, such that these small changes result in the switching of classes. Experimental results also help to validate that tidbits can assist in improving classification accuracies in real-world data sets. In this paper the consideration is assumed for analyzing cancer patient stages to detect the tidbits which is the small piece of information which leads to the other stages. This work uses classification technique used for fetching the tidbit instances to reduce the computational time.

Outlier detection is the basic step in many data mining applications. Although it is considered as a noise or an error, outliers carry some essential information. The exact definition of an outlier is based on the hidden assumption. Hawkins defines that an Outlier is an observation as to arouse suspicious object which was generated through different techniques. This can be applied by using our data mining applications based on different samples and similar data sets as shown in fig 1 which shows an outlier detection.

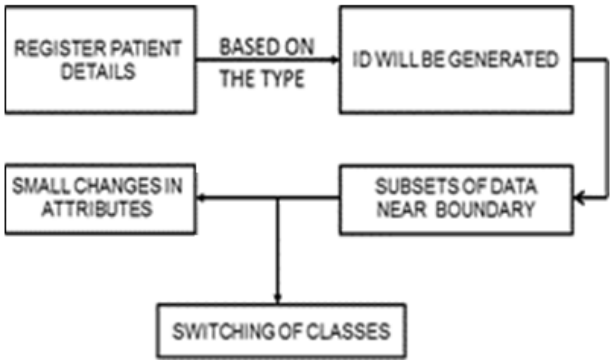


Figure 1.1: Identify tidbits

3. Defining Classifier Score

Criticality as the intrinsic worth of a subset of records. This worth is realized when the records are collectively removed from the data set or their attribute values undergo perturbation. Initial steps in defining the critical Metric (Score) relied on the effect of removing a neighborhood of data instances on a classification model. Classifier score is identified using tidbit information obtained based on the symptoms which help us to find the stage of a patient for future prediction. This idea helps us to classify different stages of patients for prevention. Result of this work shows that only subset of instances are isolated as tidbits.

A. Classifier Score using Bayes theorem

Equation (1) shows the criticality score of patient prediction based on the previous medicine and current symptoms.It can be applied for various diagnosis of cancer stages for further prediction.

$$criticalityscore = \sum {(wj)}/{n,\ j \Leftrightarrow 1...n} \cdot \quad (1)$$

Wj be the datasets
+, - be the identifiers
W j=w j+ + w j- , w j+=d j+ /d; w j- = d j-/d;
Wj lies between 0-1.
Score>wj
Then the particular stage is identified.

4. Searching Near the Class Boundary

This indicates that the potential of finding tidbits is higher along the boundary. Hence, one can focus the search along the boundary as compared to the interior. This would greatly reduce the search space as well. The stages identified are early, intermediate, and final. Based on the symptoms and datasets we search near the class boundary which includes both positive and negative data sets.

A. Detecting outliers through boundary

Searching near the boundary has the higher potential in finding the datasets that contains the false positive and true negative data items. "Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C).

A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: *http://breastscreening.cancer.gov/*."

B. Risk model data set

This dataset includes 2,392,998 screening mammograms (called the "index mammogram") from women included in the Breast Cancer Surveillance Consortium. All women did not have a previous diagnosis of breast cancer and did not have any breast imaging in the nine months preceding the index screening mammogram. However, all women had undergone previous breast mammography in the prior five years (though not in the last nine months). Cancer registry and pathology data were linked to the mammography data and incident breast cancer (invasive or ductal carcinoma in situ) within one year following the index screening mammogram was assessed.

C. Attribute Implementation

Classification of the data items using the Classification algorithms proves that EM (Expectation Maximization) is a better fit for clustering than K-Means. The existing algorithms are compared for higher accuracy and efficiency using the metric “Error Rate”. The evaluation parameters are the correctly false positive and true negative data points. Based on these parameters, the error rate is evaluated using the following formula. Per Species,

Actual Total – false positive Classified = true negative Classified
Overall Error Rate,
Error= (TN) / (TP+TN)
Here,
TN - total number of incorrectly classified species
TP - total number of correctly classified species.

5. Improving Classification Accuracy

Classification algorithms are usually judged based on the accuracy of their predictions. During the experimental stage with various data sets, tests were conducted to see if tidbits could help improve the classification accuracy.

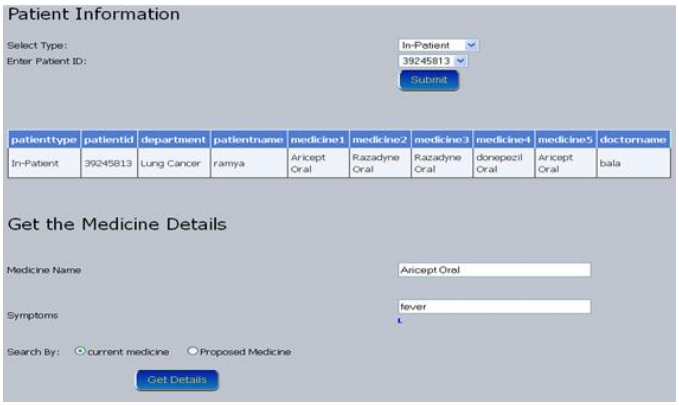


Figure 1.2: Analyzing stage

a) Accuracy

Accuracy is the percentage of obtained values compared with the expected value for the given data.The best system is that the one having highest accuracy.The following table 1.1 shows the result of accuracy obtained through possible symptoms that are identified during the classification process of tidbits. It shows accuracy of process without tidbits and with tidbits that shows the improvements in accuracy using classifier score. It also helps to identify the correct stage of a patient. The graph helps us to understand the major difference from without tidbits and with tidbits that shows the major improvements.

[1]**Table 1:** Classification Accuracy

Status	Accuracy without tidbits	Accuracy with tidbits	Accuracy increase
Early	0.046	0.25	0.20
Intermediate	0.069	0.33	0.26
Advanced	0.023	0.14	0.12

b) Implementation using WEKA tool

Weka supports several standard data mining tasks, more specifically,data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.[4] Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.as ahown in fig 1,3 shows the similarity of deviation in finding tidbits.

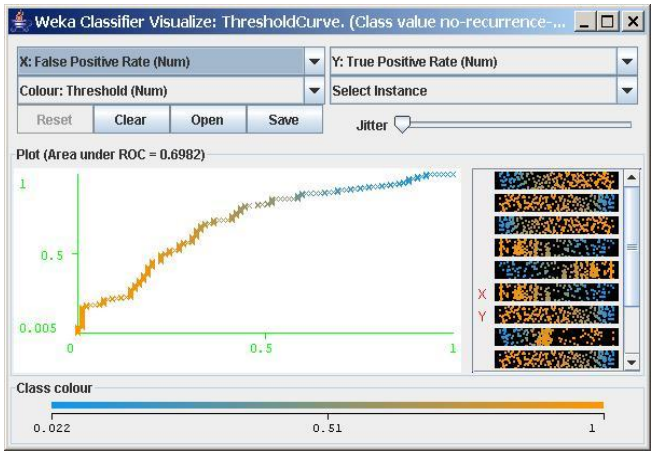


Figure 1.3: Threshold Curve

c) Naive Bayes classifier

Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier (which in typical usage are initialized with zero training instances) – if you need the Updateable Classifier functionality, use the NaiveBayesUpdateable classifier. The NaiveBayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

d) Algorithm

For example The Naive Bayes Classifier for Data Sets with Numerical AttribOne common practice to handle numerical attribute values is to assume normal distributions for numerical attributes.

The numeric weather data with summary statistics

	outlook			temperature			humidity			windy		play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$f(w) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(w-\mu)^2}{\sigma^2}}$$

Let $x_1, x_2, \ldots, x_n$ be the values of a numerical attribute in the training data set.

$$f(\text{temperatur e} = 66 \mid \text{Yes}) = \frac{1}{\sqrt{2\pi}(6.2)} e^{-\frac{(66-73)^2}{2(6.2)^2}}$$

$$= 0.0340$$

$$\text{Likelihood of yes} = \frac{2}{9} \times 0.0340 \times 0.0221 \times \frac{3}{9} \times \frac{9}{14} = 0.000036$$

$$\text{Likelihood of no} = \frac{3}{5} \times 0.0291 \times 0.038 \times \frac{3}{5} \times \frac{5}{14} = 0.000136$$

e) Steps to obtain data sets

Using this approach, Classifier model is built with training dataset And then this model is applied on test data set. Accuracy is calculated.Whenever a new patient's record is provided, GUI based Prediction System used to predict the class label. Abstractly, naive Bayes is a conditional Probability model:

f) Result and Comparison

P(A|B) = Fraction of worlds in which B is true that also have A true

P(A|B) = P(A ^ B)/P(B)

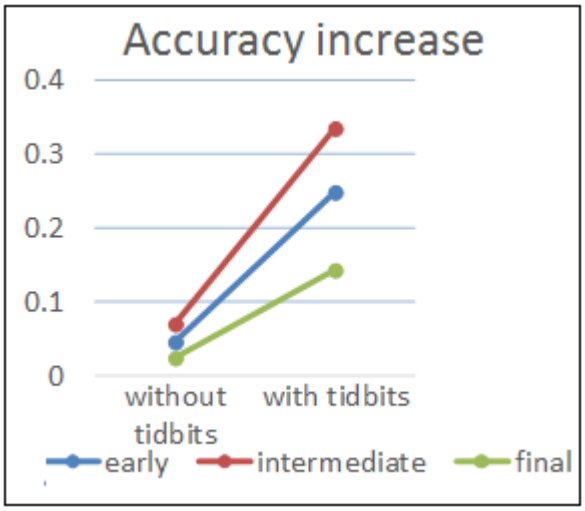
Corollary:

P(A ^ B) = P(A|B) P(B)

P(A|B)+P(¬A|B) = 1

The following graph shows the accuracy increase.

1.4 Accuracy Increase



6. Conclusion

In this paper, we discussed the implementation and efficiency details of a classifier system with similarity measures. Experiments performed on data collection using tidbits prove the efficiency of the proposed measure. This is a new technique having advantages over the other classifier systems as it can handle vague and imprecise datasets of user very well. Results indicate that proposed classifier score technique based on tidbits for handling vague, uncertain and imprecise queries. The insight provided by this model makes clear that the notions of a classifier describe situations known through imprecise, uncertain, and vague information in a way that neither replaces nor is replaced but that, rather, complements the views produced by other approaches.

References

[1] A. Bifet and E. Frank.(2010).” Sentiment knowledge discovery in Twitter streaming data, “presented at the 13th International Conference on Discovery Science, Canberra, Australia, Springer, 2010

[2] C. Ordonez and S. Pitchaimalai. Bayesian classifiers programmed in SQL. IEEE Transactions on Knowledge and Data Engineering (TKDE), 22(1):139–144, 2010.

[3] S. H. Ha and S. C. Park, “Application of data mining tools to hotel data mart on the intranet for database marketing,” Expert Syst. Application , vol. 15, no. 1, pp. 1–31, July 1998.

[4] K. W. Wong, S. Zhou, Q. Yang, M. S. Yeung, Mining customer value: from association rules to direct marketing, Data Mining and Knowledge Discovery, Issue 11, 2005, pp.57- 79.

[5] M. Maddour, M. Elloumi, A data mining approach based on machine learning techniques to classify biological sequences, KnowledgeBased Systems, Issue 15, 2002, pp.217-223.

[6] Magnus Stensmo, Terrence J. Sejnowski Automated Medical Diagnosis based on Decision Theory and Learning from Cases, World Congress on Neural Networks 1996 International Neural Network society pp. 1227-1 231 .

[7] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Survey, vol. 41, no. 3, article 15, 2009.

[8] S.D. Bay and M. Schwabacher, “Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule,”Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), L. Getoor, T.E. Senator, P. Domingos, and C. Faloutsos,eds., pp. 29-38, 2003.

[9] A. Ghoting, S. Parthasarathy, and M.E. Otey, “Fast Mining of distance-Based Outliers in High-Dimensional Datasets,” Data Mining and Knowledge Discovery, vol. 16, no. 3, pp. 349-364, 2008.

[10] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, “LOF: Identifying Density-Based Local Outliers,” SIGMOD Record, vol. 29, no.2, pp. 93-104, 2000.

[11] L. Duan, L. Xu, Y. Liu, and J. Lee, “Cluster-Based Outlier Detection,” Annals of Operations Research, vol. 168, no. 1, pp. 151- 168, http://dx.doi.org/10.1007/s10479-008-0371-9, Apr. 2009.

[12] N. Panda, E.Y. Chang, and G. Wu, “Concept Boundary Detection for Speeding Up SVMs,” Proc. 23rd Int’l Conf. Machine Learning (ICML), W.W. Cohen and A. Moore eds., vol. 148, pp. 681-688, 2006.

[13] P. Domingos, “Metacost: A General Method for Making Classifiers Cost-Sensitive,” Proc. Fifth Int’l Conf. Knowledge Discovery and Data Mining, pp. 155-164, 1999.

[14] A. Frank and A. Asuncion, “UCI Machine Learning Repository,” http://archive.ics.uci.edu/ml, 2010.

[15] R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,http://www.R-project.org, 2010.

[16] G. van Rossum et al., Python: An Object Oriented Programming Language, http://www.python.org, 1991. 1366 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 6, JUNE 2013 TABLE 3 Significance of Improvements.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, no. 1, pp. 10-18, http://doi. acm.org/10.1145/1656274.1656278,, 2009.

[18] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.[23] D. Aha and D. Kibler, "Instance-Based Learning Algorithms," Machine Learning, vol. 6, pp. 37-66, 1991. [24] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Machine Learning, vol. 95, no. 1/2, pp. 161-205, 2005.

[19] G.H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," Proc. 11th Conf. Uncertainty in Artificial Intelligence, pp. 338-345, 1995.

[20] B. Zadrozny, J. Langford, and N. Abe, "Cost-Sensitive Learning by Cost-Proportionate Example Weighting," Proc. IEEE Third Int'l Conf. Data Mining (ICDM), pp. 435-442, 2003.

[21] F. Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics Bull., vol. 1, no. 6, pp. 80-83, http://dx.doi.org/ 10.2307/3001968, 1945.

[22] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, "Anomaly Detection via Online Over sampling Principal Component Analysis", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

[23] David Sathiaraj and Evangelos Triantaphyllou, "On Identifying Critical Nuggets of Information during Classification Tasks", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 6, JUNE 2013