ISSN 2319-2720

Volume 5, No.4, October - December 2016

International Journal of Computing, Communications and Networking

Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn01542016.pdf

Attributes Collection in Medicinal Data Mining to Progress Correctness

M. N. Quadri
Assistant Professor, Dept. of Computer Science,
Nilkantrao Shinde Science & Arts College,
Bhadrawati, Dist. Chandrapur, India
mnq_1977@yahoo.com

ABSTRACT

Database size is increasing every day, data sets for analysis may include several attributes, most of which irrelevant to the mining task. Since the added volume of irrelevant or redundant attributes can measured down the mining process, dimensionality reduction reduces the data set size by removing such attributes from it. Here I employ a covering approach which leads to greater accuracy since it optimizes the evaluation measure of the algorithm while removing attributes. The best and the worst attributes are typically determined using tests of statistical significance, which assumes that the attributes are independent of one another. Data mining algorithms search for meaningful patterns in raw data sets. The Data mining process requires high computational cost when dealing with large data sets. Reducing dimensionality can effectively cut this cost. This work explains how it is often possible to reduce dimensionality with minimum loss of information. Hereby I present a method for dimension reduction applied to visual data mining in order to reduce the user cognitive load due to the density of data to be visualized and mined.

Keywords: Attributes, Data Mining, Multidimensional and Preprocessing.

1. INTRODUCTION

With the advent of high-throughput experimental technologies and of high speed internet connections, generation and transmission of large volumes of data has been automated over the last decade. As a result, science, industry, and even individuals have to face the challenge of dealing with large datasets which are too big for manual analysis. while these large “mountains” of data are easily produced nowadays, it remains difficult to automatically “mine” for valuable information within them. The quantity of stored data is almost always increasing. These data are not useful if at least a part of information they contain is not extracted. It is the goal of knowledge discovery in the databases (KDD) which can be defined as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [11]. Data mining algorithms are used for searching meaningful patterns in raw data sets. Dimensionality (i.e. the number of data set attributes or groups of attributes) constitutes a serious obstacle to the efficiency of most Data Mining algorithms. Most of the computer scientist calls this as a “curse of dimensionality”. Visual data mining is a new data mining approach using visualization as a communication channel for data mining. It lies in tightly coupling the visualizations and analytical process into one data mining tool that takes advantage of the strengths of all worlds [2, 4]. Visualization is the process of transforming information into a graphical representation allowing the user to perceive and interact with the information. Visual representation allows understanding data, determining what should be done about it. The human eye can capture complex patterns and relationships. Compared to data mining, the advantages of visual data mining are:

- The confidence in the results is improved
- The quality of the results is improved by the use of human pattern recognition capabilities
- the quality of the results is improved by the use of human pattern recognition capabilities
- If the user is the data specialist, we can use the domain knowledge during the whole process

Computer devices can display vast amount of information with various techniques. This information must be appropriately communicated to the user in order to make the best use of it. According to [5, 12], in order to be visualized, data are passed through four basic stages: independently of any visualization technique, the first step of visualization is data collection and storage. Secondly, there is a data pre-processing which goal is to transform the data into a comprehensive form. At the third step, display hardware and software are used to produce a visual representation of the data. Lastly, the users perceive, interact with the visual representation and mine it. It is necessary to address the limits of human perception. When the collected data are multidimensional, there are some limits in the third and fourth steps. For [6, 13], the conceptual boundary between low and high-dimensional data is round three to four data attributes. Their suggested guideline for characterizing dimensionality is the following: low: up to four attributes, medium: five to nine attributes and high: 10 or more. When the number of dimensions is over some dozen, the large number of axes needed to create these displays tends to overcrowd the figure, limiting the value of the plot for detecting patterns or other useful information. Our objective is to select some dimension of a data set in order to create a visualization from which relevant information can be extracted. I want to identify attributes that are significant in order to reduce dimensionality. Dimension reduction can be used to improve the efficiency of visualization of large, multidimensional data sets and may be the accuracy of algorithms used for classification in visual data mining. Knowing that:

- An optimal subset of attributes is not necessarily unique
- The visualization of more than a dozen attributes is unusable for visual data mining
- Without investigation, it is not possible to determine a dimension reduction method that can perfectly reduce the set of attributes (by taking account of different trade-offs between performance and complexity (tolerate lower performance in a model that also require less Attributes))
- The decision of a committee of experts is generally better than the decision of a single expert

2. VISUALIZATION

The role of graphics in data analysis has long been recognized as important, if not universally. In the 1930s, Fisher wrote: Diagrams prove nothing, but bring outstanding Attributes readily to the eye; they are no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them" [3, 7, 8] in order to aid our cognitive abilities. The hope is, in effect, to let the user `see' something new and from this perspective, interactivity is vital. Against this view, however, various costs must be set: Where user interaction is required, Wegman and Solka suggest that an algorithm is only feasible if it can be completed in less than a second. Without special hardware, this would apparently restrict complex analyses (of O(n2) or above) to `small' data sets; The involvement of users, even those who may be experts in a particular field, promotes subjectivity in the final result [14]; Despite the falling price of hardware, the cost of visualization systems is still considerable. It should also be realized that many effective visualizations have been created from careful consideration of the data and entirely static displays. As a result, Tufts places qualities such as design and data-ink maximization in much higher regard.

3. DIMENSIONALITY REDUCTION FOR IMPROVEMENT

If information is irrelevant or redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult. Attribute selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. Regardless of whether a learner attempts to select Attributes itself or ignores the issue, Attribute selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively. The performance of the naïve Bayes classifier is a good candidate for analyzing Attribute selection algorithms since it does not perform implicit Attribute selection like decision trees.

In this paper, I try to show that it is possible to reliably improve the naïve Bayesian classifier by applying a new Attribute selection algorithm that is both simple and effective.

3.1 Naïve Bayes and NB classifier

Naïve Bayes, a special form of Bayesian network has been widely used for data classification. As a classifier it learns from training data from the conditional probability of each attribute given the class label. Using Bayes rule to compute the probability of the classes given the particular instance of the attributes, prediction of the class is done by identifying the class with the highest posterior probability. Computation is made possible by making the assumption that all attributes are conditionally independent given the value of the class.

Naïve Bayes is best understood from the perspective of Bayesian networks. Bayesian networks (BN) graphically represent the joint probability distribution of a set of random variables. A BN is an annotated directed acyclic graph that encodes a joint probability distribution over a set of attributes X. Formally a BN for X is a pair B= <G,Ɵ> , where G represents the directed acyclic graph whose nodes represent the attributes X1, X2,..Xn and whose edges represent direct dependencies between the attributes. The BN can be used to compute the conditional probability of a node given values assigned to the other nodes. The BN can be used as a classifier where the learner attempts to construct a classifier from a given set of training examples with class labels. Here nodes represent dataset attributes. Assuming that X1, X2,..Xn are the n attributes corresponding to the nodes of the BN and say an example E is represented by a vector x1, x2,..xn where x1 is the value of the attribute X1. Let C represent the class variable and c its value corresponding to the class node in the Bayesian network, then the class c of the example E (c(E)) can be represented as a classifier by the BN [Harry Zhang, Charles X. Ling,] as

$$c(E)= \underset{c \in C}{\arg\max}\ p(c)\ p(\ x_1, x_2, \ldots. x_n \mid c) \qquad (1)$$

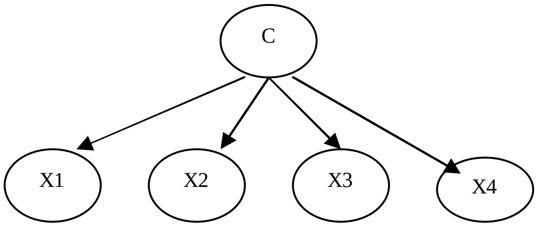

Figure 1: Example Structure of Naïve Bayes

Although Bayesian networks can represent arbitrary dependencies it is intractable to learn it from data. Hence learning restricted structures such as naïve Bayes is more practical. The naïve Bayesian classifier represented as a BN has the simplest structure. Here the assumption made is that all attributes are independent given the class and equation 1 takes the form

$$c(E) = \underset{c \in C}{\arg\max}\ p(c) \prod_{i=1}^{n} p(x_i \mid c) \qquad (2)$$

The structure of naïve Bayes is graphically shown in Fig.1. Accordingly each attribute has a class node as its parent only [5]. The most likely class of a test example can be easily estimated and surprisingly effective. Comparing naïve Bayes to Bayesian networks, a much more powerful and flexible representation of probabilistic dependence generally did not lead to improvements in accuracy and in some cases reduced accuracy for some domains.

3.2 Attribute Selection for NB Classifier

Attribute selection is often an essential data preprocessing step prior to applying a classification algorithm such as naïve Bayes. As a learning scheme naïve Bayes is simple, very robust with noisy data and easily implementable. I have chosen to analyze Attribute selection algorithms with respect to naïve Bayes method since it does not perform implicit Attribute selection like decision trees.

Algorithms that perform Attribute selection as a preprocessing step prior to learning can generally be placed into one of two broad categories. One approach referred to as the ‘wrapper’ employs as a subroutine a statistical re-sampling technique such as cross validation using the actual target learning algorithm to estimate the accuracy of Attribute subsets. This approach has proved useful but is slow because the learning algorithm is called repeatedly. The other approach called the ‘filter’ operates independently of any learning algorithm. Undesirable Attributes are filtered out of the data before induction commences. Although filters are suitable to large datasets they have not proved as effective as wrappers. Generally the filter approach computationally more efficient than the wrapper approach. The wrapper approach, on the other hand involves the computational overhead of evaluating candidate Attribute subsets by executing a selected learning algorithm on the dataset represented using each Attribute subset under consideration. Hence I try to propose a new algorithm for Attribute selection.

4. NEW DIMENSION REDUCTION ALGORITHM

To obtain the best accuracy in attribute selection, I try to propose to use random search and (backward, forward ((like sequential floating selection), knowing that the function used is non monotonic [1])).

The wrapper approach allows rising to interesting details for the data analysis specialist (data mining domain). Knowing that the classifier error rate capture two basic performance aspects: class separate ability and any structural error imposed by the form of the classifier. Other types of details, namely, properties that good dimension sets are presumed to have (class separate ability or a high correlation between the attributes) are more appropriate to statistician. These details could not be highlighted at all by the wrapper methods. In order to take this fact into consideration, I have added some filter-based criteria (consistency, entropy, distance) to our attribute subset selection method. In input, there is a data set and the output is a subset of attributes of this data set. The generation procedure uses a combination of random search and sequential floating selection. Concerning the evaluation functions, I try to use a combination of filter (consistency, entropy, distance) and wrapper ((LDA, QDA, KNN) [Ripley, 1996]). LDA, QDA, KNN executions use ten fold cross validation. At each step of the execution of these algorithms, the following evaluation criteria are used: the correctness of the classification rule, the accuracy, the ability to separate classes, and the confidence. Next, I have combined their selected attribute subset in order to derive a consensus of the most suitable subset of attributes. For this purpose, a learning step, based on the results of generation procedures evaluated by filter-based criteria and wrapper based approaches enables us to lead to final results.

More precisely, the domain I have consider consist of a set of N = 6 experts (consistency, entropy, distance, LDA, QDA, KNN evaluation functions) E = {e1, ..., e_N}, a set of dimension subsets DS = {D1, ...,D_K}, where K is not a constant. Attribute subsets are available for expert/subset pairs {e,D}, where e ∈ E and D ∈ DS. I define preference of a dimension d as the probability that the dimension appears in the experts feature subsets, p(d) = $\sum p_i(d)$. P_i(d) represents the probability that expert i selects dimension d . pi(d) = y /Z if expert i has selected featured, 0 otherwise. y is the number of selected dimensions. Z represents the number of attributes in the original data set. The preference value of features is used in order to pool together the selected features and to rank them. Next, if the pool number of dimensions is greater than twenty (number of attributes which can be correctly display and visually mine), it is divided into relevant attributes (consensus) and less relevant attributes. At the cutting point, if some features have the same preference value, I have use expert relevance score (ERS) in order to determine which features match the best. For each feature in the conflicting part, the decision to add it in consensus part of the pool or not is made according to the relevance score of the experts who choose the feature. The selected features are those with great expert relevance score computed as following:

ERS = g /T , where g represents the number of attributes in the consensus part which have been selected by the expert and T the total number of features selected by that expert.

5. EXPERIMENTS AND RESULTS

In order to test proposed approach, I have compared its results with the results of two widely used attribute selection methods. Namely, R language implementations of: Las Vegas Filter [14] (package dprep) and a wrapper based feature selection algorithm (Stepclass, package klaR). Our consensus based algorithm is also implemented in R. I have use a desktop Intel centrino , processor 1.70 GHz, Windows to perform these tests. The data sets (from the UCI [9] and the Kent Rigde Bio-Medical Data Set repositories [10] were chosen because of their large number of attributes (Table 1).

Table 1: Data set explanation

Name	NbAt	NbInst	NbClass
Lung Cancer	57	32	3
Promoter	59	106	2
Sonar	60	208	2
Arrhythmia	280	452	16
Isolet	618	1560	26
ColonTumor	2000	62	2
CentralNervSyst	7129	60	2

The final results of LVF, stepclass and consensus based algorithm were evaluated by IBk, a K nearest neighbor algorithm (KNN) found in WEKA, a free Java-based, open source, that provide a variety of machine learning algorithms.

Table 2 shows the difference (attribute size and KNN accuracy) between the original and the final data sets. The attribute subset selected by the consensus based approach (less or equal to 20) allows visualizing and mining the whole data sets. The change in the accuracies of KNN classifier is minimal or there is no change. This is not the case of LVF or step class (table 3). The data set Arrhythmia for example has a subset with 109 attributes (LVF results) and for the data set Promoter, stepclass does not reduce the dimension.

Table 2: Evaluation of Number of Attributes and Accuracy with Knn Algorithm before and after Reduction

Name	Initial NbAt	Final NbAt	Accuracy Before	Accuracy After
Lung Cancer	57	4	37.50%	75.00%
Promoter	59	9	85.84%	68.87%
Sonar	60	8	86.54%	71.15%
Arrhythmia	280	4	53.44%	59.96%
Isolet	618	14	85.57%	70.24%
Colon Tumor	2000	19	77.42%	79.03%
CentralNervSyst	7129	20	56.67%	60.00%

Our goal is firstly to reduce the number of dimensions in order that the data set could be visualized. Table 3 shows that I attend decided principal goal and I have obtained results that are comparable to those of the attribute selection algorithms which objective is to improve classifiers accuracy.

Indeed, the consensus based approach allows obtaining the best result for data set Lung-Cancer and about the same accuracy rate for the data sets Sonar, Arrhythmia and Colon Tumor. It should be noted that two cases arise: either the attributes of the data set to be treated are redundant or irrelevant and then the results are comparable with those of filters or wrappers based approaches or it does not exist redundancy in the attributes and dimension reduction implies a loss of accuracy.

The data sets in this category are: Isolet (best accuracy with LVF for 268 attributes) and Promoter (best accuracy with Stepclass for 59 attributes). For these data sets, the number of selected dimensions in spite of the best accuracy remains unusable for visual data mining.

Table 3: Comparison of number of attributes and accuracy with KNN algorithm before and after reduction

Name	Final NbAt	LvF NbAt	Wrappp NbAt	Final Acc	LvF Acc	Wrapp Acc
Lung Cancer	4	17	4	75.00%	62.50%	71.87%
Promoter	9	16	59	68.87%	80.19%	85.85%
Sonar	8	18	4	71.15%	82.21%	71.63%
Arrhythmia	4	109	4	59.95%	54.65%	60.84%
Isolet	14	268	8	70.24%	83.00%	57.98%
Colon Tumor	19	918	5	79.03%	77.42%	79.03%
CentralNerv Syst	20	3431	8	60.00%	58.33%	71.67%

6. CONCLUSION

In this research work an attempt was made to evaluate feature selection with naïve Bayes classifier that could be used for medical data mining.The data visualization, the performance of classification algorithms are affected by attributes. When a data set has a large number of attributes, it is impossible to perform visual data mining. Irrelevant, redundant features have a negative effect on the accuracy of a classifier and on visual representations. I have defined a dimension reduction method for visual data mining. Then I have compared successfully the results of this framework to two widely used attribute selection algorithms.

REFERENCES

[1] P. Pudil, J. Novovicova and J. Kittler, Floating search meathods in feature selection Pattern Recognition Letters, pp. 1119–1125, 1994.

[2] P.C. Wong, Visual data mining, IEEE Computer Graphics and Applications, pp. 20–21, 1999.

[3] D.W. Scott, Multivariate Density Estimation: Theory, Practice, Visualization, Wiley Series in Probability and Mathematical Statistics Wiley.

[4] C. Silva, E. Groeller, and H. Rushmeier, The value of visualization, In editors, Proceedings of IEEE Visualization 2005, 2005.

[5] H. Zhang and C. X. Ling, "A Fundamental Issue of Naïve Bayes", In proceedings of the *Canadian Conference on Artificial Intelligence*:, pp. 591-595, 2003.

[6] B. Cestnik, "Estimating probabilities: A crucial task in machine learning", In Proceedings of the 9th European Conference on Artificial Intelligence, pp. 147–149. 1990.

[7] H. Liu, R. Sentino, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.

[8] Ranjit Abraham, B. Simha, S. S. Iyengar, "A comparative study of discretization and feature selection methods for Medical datamining using Naïve Bayesian classifier", Indian Conference on Computational Intelligence and Information Security (ICCIIS '07), pp.196-201, 2007.

[9] C. Blake and C. Merz. UCI Repository of machine learning databases. Irvine, University of California, Department of Information and Computer Science, from www.ics.uci.edu/˜ mlearn/MLRepository.html, 1998

[10] L. Jinyan and L. Huiqing. Kent Ridge Bio-medical Data Set Repository. http://sdmc.lit.org.sg/GEDatasets, 2002.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, and G. Smyth. Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press, Menlo Park, CA, 1996

[12] C. Ware. Information visualization, Perception for design. Morgan Kaufman Publishers, San Diego, USA, 2000.

[13] O. Ferreira and Levkowitz. From visual data exploration to visual data mining: a survey, visualization and computer graphics. IEEE Transactions, pages 378–394, 2003.

[14] J.J. van Wijk, The value of visualization. In C. Silva, E. Groeller, and H. Rushmeier, editors, Proceedings of IEEE Visualization 2005