www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 11 Nov 2015, Page No. 15091-15094

Study of Classification of Diseases by Genetic Algorithm for Multiclass Support Vector Machine Using Hadoop

Mr.Ankit R. Deshmukh [1], Prof.S.P.Akarte [2]

[1] M.E. (CSE), Second Year, Dept. of Computer Science, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati

Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

ankitdeshmukh100@gmail.com

[2] Assistant Professor, Dept. of Computer Science, Prof. Ram Meghe Institute Of Technology and Research, Badnera Amravati.

Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

s_akarte25@rediffmail.com

Abstract: Since many years ago, the scientific community is concerned about how to increase the accuracy of different classification methods, and major achievements have been made so far. Hadoop and MapReduce are used to handle these large volumes of variable size data. This work focuses on the combining a feature selection technique based on genetic algorithm and support vector machines (SVM) of medical disease classification. SVM is relatively a novel classification technique and has been shown higher performance than traditional learning methods in many applications. The idea is to use GA as an optimizer to find the optimal values of hyper-parameters of SVM and adopt a supervised learning approach to train the SVM model. In this paper we propose a genetic algorithm (GA) based classification method.

Keywords: Classification, Genetic Algorithm, Hadoop, Map-Reduce, Support Vector Machines.

1. Introduction

Classification for imbalance data sets has been studied by machine learning community since the last decade [1]. The under sampling and over sampling method [2] balances the data sets by randomly selecting small number of objects from majority class, and doubling the objects in the minority class. The main drawback is some import points, such as support vectors, may be neglected by the random algorithm.[3] pointed out the under sampling strategy is not a good choice for SVM, and the over sampling cannot improve the final accuracy.

The classification of medical data has become an increasingly challenging problem, due to recent advances in medical mining technology. Classification of this tremendous amount of data is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications. In this work, we develop an approach to optimize the support vector machine parameters which combines the merits of support vector machine (SVM) and genetic algorithm (GA).

The remainder of this paper is organized as follows. Section III provides Hadoop Overview and Section IV describes the Hadoop Map Reducer, Section V describes Multiclass Classification, Section VI gives idea about Genetic Algorithm and Section VII use of Support Vector Machine, and finally section VIII concludes this paper.

2. Literature Review

There are many research works that try to improve traditional techniques or develop new algorithms to solve the class imbalance problem. However, most of those studies are focused only on binary case or two classes. Only a few researches have been done for multiclass imbalance problem that is much more common and complex in the real-world application. We will study the multiclass imbalanced data problem, and developed new classification algorithms that can effectively handle the imbalance problem in many biomedical domains. Crammer K, Singer Y, Cristianini N, Shawe-taylor J,Williamson B. Implemented the algorithm of multiclass kernel-based vector machines. Wasikowski M, Chen X W. worked on the small sample class imbalance problem using feature selection. Similarly, Chen X, Gerlach B, Casasent D. introduced support vectors for imbalanced data classification.

Imbalanced data is a common and serious problem in many biomedical classification tasks. It creates lots of confusion on the training of classifiers and results in lower accuracy of minority classes prediction [13]. This is engendered due to the less availability or by limitations on data collection process such as high cost or privacy problems. The majority classes' overloads standard machine learning algorithms that it cannot bear the load and traditional classifiers making confusions between the decisions towards the majority class and try to optimize overall accuracy. To improve traditional methods or to develop new algorithms for solving the problem of class

imbalance, many researches were done. Most of those studies are focused only on binary case or two classes. Only a few researches have been done for multiclass imbalance problem which are more common and complicated in structure in the real-world application.

3. Hadoop Overview

Hadoop is a distributed computing framework released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and it can be efficient, reliable, scalable way to process data. Its core idea is to build on a large number of cheap and efficient cluster hardware devices, in the form of software processing to provide storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems.

Hadoop is a Map Reduce programming model and mass data. It has made a lot of simulation system in the cloud computing, such a calculation based on the concept of cloud modeling and simulation platform of COSIM-CSP system, a new mode of the networked manufacturing, private cloud framework for visual simulation, and the military training system.

4. Hadoop Map Reducer

The term Hadoop [16] comprises a family of many related projects with the same infrastructure for distributed computing and large-scale data processing. It is better known for the Map Reduce algorithm, shown below, and its distributed file system HDFS, which runs on large clusters of commodity machines. Hadoop was created by Doug Cutting and has its origins in Apache Nuts, an open source web search engine. In January 2008 Hadoop was made a top-level project at Apache, attracting to itself a large active community, including Yahoo!, Facebook and The New York Times. At present, Hadoop is a solid and valid presence in the world of cloud computing.

Map Reduce is a programming model whose origins lie in the old functional programming. It was adapted by Google as a system for building search indexes, distributed computing and database communities. It was written in C++ language and was made as a framework, in order to simply develop its applications. In Hadoop programs are mainly in Java language but it is also possible, through a mechanism called "streaming", to develop programs in any language that supports the standard I/O. Map Reduce is a batch query processor and the entire dataset is processed for each query. It is a linearly scalable programming model where users programs at least two functions: the "map" function and "reduction" functions. These functions process the data in terms of key/value pairs which are unaware of the size of the data or the cloud that they are operating on, so they can be used unchanged either for a small dataset or for a massive one.

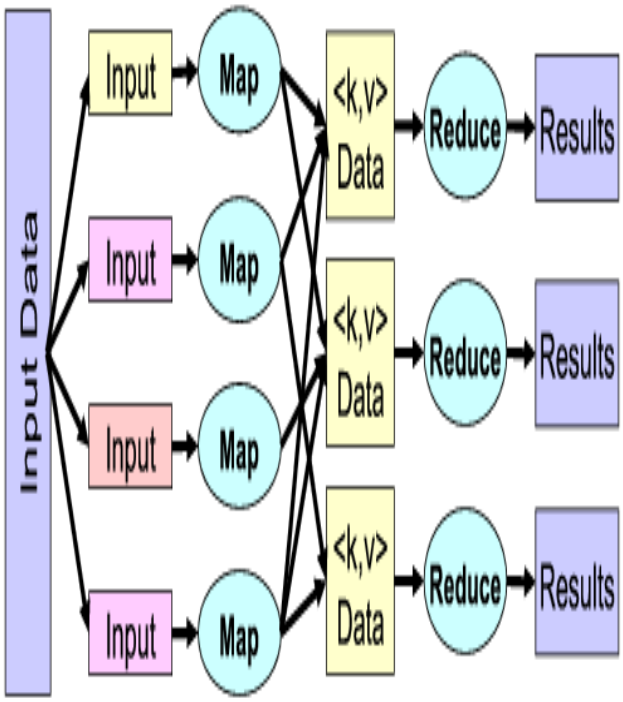

Figure 1: Mapping & Reduction

5. Multiclass Classification

Multiclass or multinomial classification is the problem of classifying instances into more than two classes. While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies. Multiclass classification should not be confused with multi-label classification, where multiple labels are to be predicted for each instance [14].

A. Imbalanced Data

Errors of major class instances will control the total error. Thus, a classifier will certainly be biased toward the major class to minimize the total errors, as shown in the figure

B. Noisy Data

An error of each point can take values ranging from 0 to $+\infty$, therefore errors of a few bad or noisy points can critically compromise the overall errors which result in impairment of classifier's performance. From both cases that the summation of all errors is not suitable for use as objective function of the optimization problem [15].

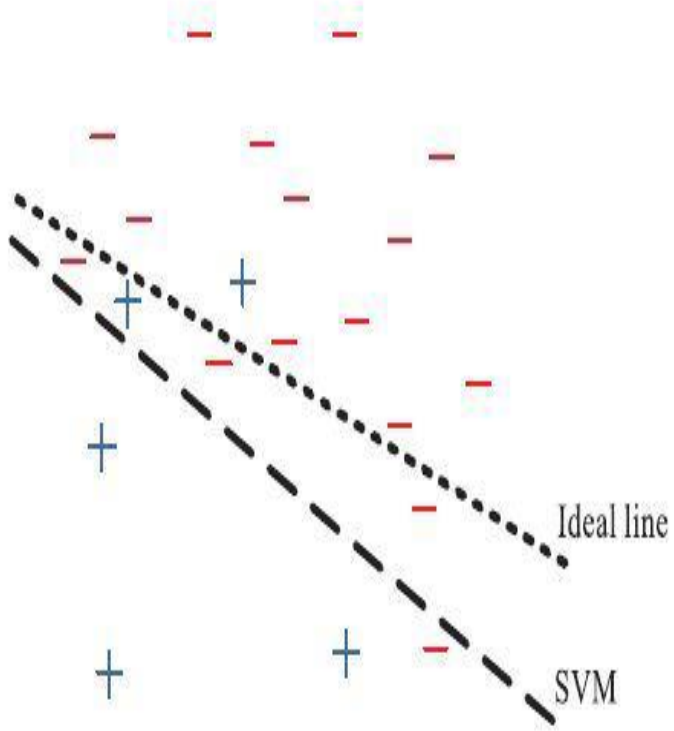

Figure 2: SVM on imbalanced dataset is biased toward the major class**.**

6. Genetic Algorithm

The Genetic Algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. Generally GAs are not used to find patterns, but rather to guide the learning process of data mining algorithms such as neural nets. A GA allows a population composed of many individuals [4] (basically the candidates) to evolve under specified selection rules to a state that maximizes the fitness. GA is known as a subset of evolutionary algorithms that model biological processes which is influenced by the environmental factor to solve various numerical optimization problems. GA allows a population composed of many individuals or called chromosomes to evolve under specified rules to a state that maximizes the fitness or minimizes the cost functions. A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation; crossover combines good strings to generate better offspring; mutation alters a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. The GA cycle continues until the termination criterion is reached. In feature selection, Genetic Algorithm (GA) is used as a random selection algorithm, Capable of effectively exploring large search spaces [5].

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems [6].

7. Support Vector Machine

Support Vector Machine (SVM) is inspired on statistical learning theory developed by Vapnik on 70's [7]. It achieves optimal classification in linear separable case. It is better than neural networks [8], decision trees [9] and Bayesian classifiers [10] in some applications. SVM offers a hyperplane that represents the largest separation (or margin) between two classes [11]. This kind of maximum-margin hyperplane may not exist because of class overlapping or mislabeled examples. The soft margins SVM by introducing slack variables can find a hyperplane that splits the examples as cleanly as possible. However, SVM requires balance data and it does not consider the classes' distribution.

Support Vector Machine (SVM) is a classification technique based on statistical learning theory. It is based on the idea of a hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized [12].The figure below presents an overview of the SVM process

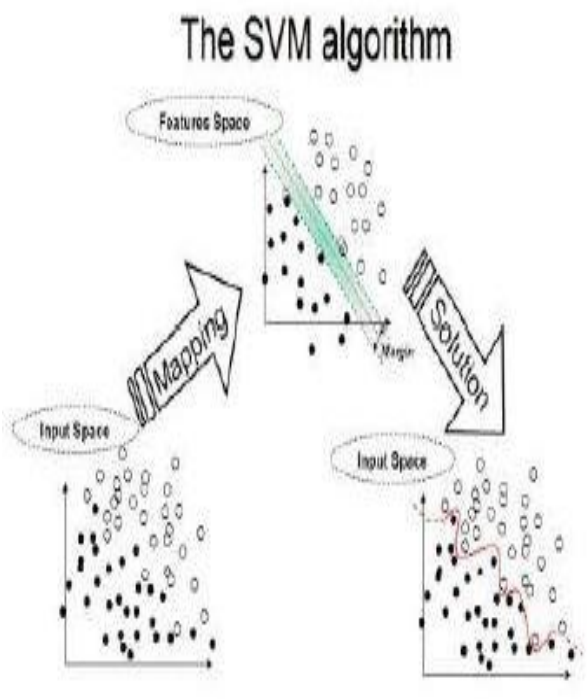

Figure 3: SVM Process

8. Conclusion

This paper gives the idea to design Support Vector Machine and Genetic Algorithm was analyzed to find the classification accuracy and also proposed a Genetic algorithm based optimization algorithm, which can optimize the parameter values for SVM, and obtain the optimal subset of features.

References

[1] Z.-Q. Zeng and J. Gao, "Improving svm classi_cation with imbalance data set," in *Proceedings of the 16th International Conference on Neural Information Processing*, Springer-Verlag, 2009, pp.389–398.

[2] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," *In Proc. ECML*, 2004, pp.39-50.

[3] Suzan Koknar-tezel and Longin Jan Latecki,Improving SVM Classi_cation on Imbalanced Data Sets in Distance Spaces, *IEEE International Conference on Data Mining*, 2009, pp 259 - 267.

[4]. Jihoon Yang and Vasant Honavar. Feature subset selection using Genetic Algorithm. IEEE Intelligent Systems, 1998.

[5]. L. Chu, and C. Wu, "A Fuzzy Support Vector Machine Based on Geometric Model," Proceedings of the fifth World Congress on Intelligent Control and Automation, Hangzhou, P.R. China, pp.1843-1846, June 15-19, 2004.

[6] Mitchell, Melanie (1996). An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press. ISBN 9780585030944.

[7] Vapnik V., "The Nature of Statistical Learning Theory," *Springer, N.Y.*, 1995.

[8] Ra¸sit Köker, A genetic algorithm approach to a neural-network-based inverse kinematics solution of robotic manipulators based on error minimization, Information Sciences, Volume 222, 10 February 2013, Pages 528-543.

[9] J. Chen, C. Wang, and R. Wang, Combining support vector machines with a pairwise decision tree, *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 409 –413, july 2008.

[10] J. Cervantes, X. Li, and W. Yu, "Splice site detection in dna sequences using a fast classi_cation algorithm," 2009 *IEEE international conference on Systems, Man and Cybernetics*, Piscataway, NJ, USA, 2009, pp. 2683–2688.

[11] N.Cristianini, J.Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* , Cambridge University Press, 2000.

[12] C.J.C. Burges, " A tutorial on support vector machines for pattern recognitation " , Data Mining and Knowledge Discovery , vol.2 , 1998, pp. 121-167.

[13] Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004,6: 1-6.

[14] Ferri C, Hernndez-orallo J, Salido M. Volume under the ROC surface for multi-class problems. exact computation and evaluation of approximations. In: Proc. of 14th European Conference on Machine Learning. Cavtat- Dubrovnik, Croatia, 2003: 108-120.

[15] Yang X Y, Liu J, Zhang M Q, Niu K. A new multi-class SVM algorithm based on one-class SVM. In: Proceedings of the 7th International Conference on Computational Science. Beijing, China, 2007: 677-684.

[16] T. White, *Hadoop: The Definitive Guide*, Third. O'Reiily, 2012.

Author Profile

Ankit R. Deshmukh, ME (CSE) ,Second Year, Department of CSE,Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

Prof. Shrikant P. Akarte, Assistant Professor,Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati. Sant Gadgebaba Amravati University, Amarvati, Maharashtra, India – 444701.