International Journal of Technology and Engineering System (IJTES)
Vol 7. No.4 2015 Pp. 394-397
©gopalax Journals, Singapore
available at : www.ijcns.com
ISSN: 0976-1345

A COMPARATIVE STUDY ON MULTI-TARGET SOLAR-FLARE DATASET

N.SHEELARANI
Student of Computer Science and Engineering,
University College of Engineering, Anna University::Regional Centre,
Tiruchirappalli, Tamilnadu, India
snaga4918@gmail.com

ABSTRACT

The chief aspect to determine the multi-label classification from a usual classification task is that a number of class variable values have to be predicted contemporaneous. Multi-label classification is a type of supervised learning where the classifier is obliged to learn from a set of example; each example can belong to multiple classes and so after be able to predict a set of class variable values for a new instance. There survive a broad scope of uses for multi-labeled predictions such as gene functionality classification, semantic image labeling, text categorization, medical diagnosis etc. This study paper presents the performance of different base classifier for the multi-dimensional solar-flare dataset with two multi-label classifier, performance of the classifier is evaluated and the metrics used are hamming loss, exact match, hamming score, accuracy and total time.

Keywords: multi-dimension, multi-label and multi-label classifier.

1. INTRODUCTION

Information and data are being accumulation is increasing never seen in traditional methods of handling those huge amounts are not sufficient. Understanding the relationship in the huge volume of data is critical for a variety of problems ranging from determining what procedures are most effective to how best to categorize the different types of solar flare in a period of diminishing resources [1]. One popular approach that is frequently used and quite efficient in analyzing data is Data Mining. Today, data mining plays a vital role in widely used to understand market patterns, customer behavior, fault detection and fraud detection etc. Data mining can be applied to different tasks related to decision-making. The major challenges in data mining are huge volume of data, regular update, inconsistent data representation, and poor integration, noise, number of variables and missing or incomplete. Due to huge volume of information, field knowledge can be used to remove unwanted records in decreasing the size of the database. Data mining techniques would be less sensitive to noise. If the number of variable increases, then the computational complexity is not elongate for certain data mining technique.

Various data mining techniques such as classification, clustering, regression and association rule mining are used in data mining applications. Data mining algorithms are appropriately used for capable of improving the quality of prediction, diagnosis and text categorization. The three target attributes of solar flare correspond to types of solar-flare seen in a 24 hour period. Single classification task deal with trouble where each item should be assigned to exactly one category from a finite set of available labels. In multi-label classification where each item is associated with multiple class variables contemporaneous. There exists a broad range of applications in multi-label classification in everyday life. In recent years it has received more attention from the machine learning community and many recent studies look for efficient and accurate algorithms for this challenge [2].

An exhaustive approaches for multi-label classification and partition them into two main categories. (1) Problem transformation- methods that transform the multi-label classification problem into one or more single label

classification problems. (2). Algorithm adaptation-methods that extend specific learning algorithm in order to handle multi-label data directly. The common methods available in the problem transformation are Binary Relevance, Labels Power-set and classifier chain approaches. In Binary relevance approach, a multi-label classification problem is decomposed into multiple, independent binary classification problems, and the final labels for each data point are determined by aggregating the classification results from all binary classifiers.
The rest of the paper is structured as follows: related work is discussed in the next section. In section 3, materials and method is discussed. Section 4 presents the result analysis and section 5 presents the conclusion.

2. RELATED WORK

Multi-label classification is a special supervised learning issue. Different comparisons of multi-label learning approaches are made. The data scattered problem of the LP approach was addressed. Another approach for multi-label classification in domains with a large number of labels was proposed. Read et al., argues in defense of the binary relevance method. They present a method for chaining binary classifiers- Classifiers Chains in a way that overcome the label independence assumption of binary relevance.

Methods for multi-label learning are binary relevance method, pair-wise method, label power-set method, algorithm adaptation method and ensemble methods [3]. Binary relevance is the well-known one-against-all strategy. It addresses the multi-label learning problem by learning one classifier for each label. The pair-wise method consists of two techniques. They are calibrated label ranking and quick weighted voting method. Calibrated label ranking is a technique for extending the common pair wise approach to multi-label learning. It introduces an artificial label, which represents the split-point between relevant and irrelevant labels. The quick weighted voting method for multi-class classification is a variant of the calibrated label ranking method that introduces a more effective voting strategy than the majority voting used by the calibrated label ranking method. Hierarchy of multi-label classifiers is an algorithm for effective and computationally efficient multi-label learning in domains with a large number of labels [3].

Algorithm Adaptation method is a method to adapt directly to the multi-label classification task. In this method, multi-label C4.5 is an adaptation of the well known C4.5 algorithm for multi-label learning by allowing multiple labels in the leaves of the tree. Predictive clustering trees are decision trees viewed as a hierarchy of clusters. Multi-label k-Nearest neighbors is also an algorithm adaptation method. in this method, for each test instance, its k-nearest neighbors in the training set are identified [3] Ensemble method consists of four techniques such as Random k- label sets, Ensembles of classifier chain. RAkEL is an ensemble method for multi-label classification. It draws m random subsets of labels with size k from all labels and trains a label power-set classifier using each set of labels. Ensembles of classifier chain are an ensemble multi-label classification technique that uses classifier chains as a base classifier [3].

3. MATERIAL AND METHODS

We collected a dataset from the MEKA website. The

Instances	c	K	Nominal attributes	Label cardinality train/test	Train /Test	Label density
323	3	5	10	0.268/0.336	213/110	0.101

solar-flare dataset contains 323 examples, 3 target variables and each target variable has 5 class labels and 10 nominal attributes. A solar flare is a sudden flash of brightness observed over the Sun's surface or the solar limb, which is interpreted as a large energy release of up to 6×10^{25} joules of energy. Solar flares are classified as A, B, C, M or X according to the peak flux (in watts per square metre, W/m^2) of 100 to 800 Pico metre X-rays near Earth.

Classification	Range
A	$<10^{-7}$
B	10^{-7}-10^{-6}
C	10^{-6}-10^{-5}
M	10^{-5}-10^{-4}
X	10^{-4}-10^{-3}
Z	$>10^{-3}$

Table 1: Types of Solar-flare

Attributes are Code for class (modified Zurich class) -(A,B,C,D,E,F,H), Code for largest spot size-(X,R,S,A,H,K), Code for spot distribution-(X,O,I,C), Activity-(1 = reduced, 2 = unchanged),Evolution- (1 = decay, 2 = no growth,3 = growth), Previous 24 hour flare activity code-(1 = nothing as big as an M1,2 = one M1,3 = more activity than one M1),Historically-complex-(1 = Yes, 2 = No), Did region become historically complex (1 = yes, 2 = no)on this pass across the sun's disk, Area-(1 = small, 2 = large),Area of the largest spot(1 = <=5, 2 = >5).

Table 2: Statistics of Solar Flare Dataset

From the table, we depicted the number of instances in the Solar Flare dataset is 323 and number of class variables is 3 and the number of labels in each class variables is 5 and the

10 attributes are nominal, Label cardinality of train and test is 0.268 and 0.336, number of train and test instances are 213 and 110 and label density is 0.101.
Methods used in this paper are J48, kNN, and Naive Bayes are base classifier and Binary relevance and classifier chain are used as multi-label classifier.

Multi-label classifier	Base classifier	Hamming score	Exact match	Hamming loss	Accuracy	Total time
Binary Relevance	J48	0.912	0.791	0.088	0.912	0.028
	kNN	0.9	0.755	0.1	0.9	0.03
	Naive Bayes	0.858	0.727	0.142	0.857	0.009
Classifier Chain	J48	0.912	0.791	0.088	0.912	0.021
	kNN	0.9	0.755	0.1	0.9	0.014
	Naive Bayes	0.845	0.718	0.155	0.845	0.012

4. EXPERIMENTAL ANALYSIS

Classifier such as Base classifier and multi-labelled classifiers and five different metrics has been involved in the experiment.
The following table predicted the performance metric of the two multi-label classifiers and 3 base classifier. The performance metrics are hamming score, exact match, hamming loss, accuracy and total time. Execution time is one of the performance metrics. From the table 3 classifier chain execution time is less. Based on the exact match metric the J48 base classifier performance is best.

Table 3: Performance of Classifiers

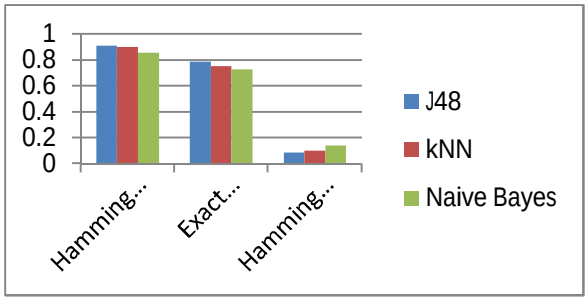

Fig 1: Performance of Binary relevance Classifier.

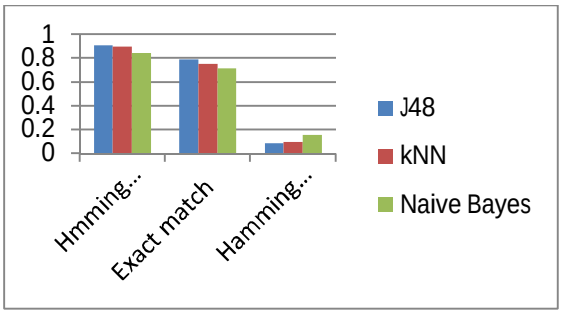

Fig 2: Performance of Classifier Chain Classifier.

Figure 1 represents the graphical representation of the performance of Binary Relevance classifier and figure 2 represents the graphical representation of the performance of Classifier Chain classifier.

5. CONCLUSION

In this paper we performed base classifier of J48, kNN, Naïve bayes and used two Multi-Label classifiers such as Binary Relevance and Classifier Chain. Execution time is one of the performance metrics. From the table 3 classifier chain execution time is less. Based on the exact match metric the J48 base classifier performance is best. In future, we can apply multi-target classifiers.

REFERENCES

[1]. Soman K.P., Diwakar Shyam and Ajay v. "Insight into Data Mining: Theory and practice", PHI 2009.
[2]. L.Tenenboim, L.Rokach, and B.Shapria, "Indentification of label dependencies for multi-label classification", in proc. 2nd Int.Workshop Learn. MLD ICML/COLT, Haifa, Israel, 2010.
[3]. G.Madjarov, D.Kocev, D.Gjorgjevikj, and S.Dzeroski, " an extensive experimental comparison of methods for multi-label learning", Pattern Recognition, vol.45, no.9, pp.30841-3104, 2012.
[4]. A. Andrew. (2010).Frank and Arthur Asuncion. UCI machine learning repository[Online].Available: http://archive.ics.uci.edu/ml.
[5]. A. de Carvalho and A. A. Freitas, 2009. “A tutorial on multi-label classification techniques,”in Studies in Computational Intelligence 205, A. Abraham, A. E. Hassanien, and V. Snásel, Eds. Berlin, Germany: Springer, pp. 177–195.
[6]. C. Bielza, G. Li, and P. Larrañaga, 2011. “Multi-dimensional classification with Bayesian networks,” Int. J. Approx. Reason., vol. 52, no. 6, pp. 705–727.
[7].David J. Hand, Heikki Mannila and Padhraic Smyth, 2005. “Principles of Data Mining”, (Adaptive Computation and Machine Learning).

[8].G. Tsoumakas, I. Katakis, and I. P. Vlahavas, 2010. "Mining multilabel data," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Berlin, Germany: Springer.
[9].G. Tsoumakas and I. P. Vlahavas, 2007. "Random k-labelsets: An ensemble method for multilabel classification," in Proc. 18th ECML,Warsaw, Poland, pp. 406–417.
[10].J. Read, B. Pfahringer, and G. Holmes, 2008. "Multi-label classification using ensembles of pruned sets," in Proc. 8th IEEE ICDM, Pisa, Italy, pp. 995–1000.
[11].J. Read, B. Pfahringer, G. Holmes, and E. Frank, 2009. "Classifier chains for multi-label classification," in Proc. 20th ECML, Bled, Slovenia, pp. 254–269.
[12].J. Read, B. Pfahringer, G. Holmes, and E. Frank, 2011. "Classifier chains for multi-label classification," Mach. Learn., vol. 85, no. 3, pp. 333–359.
[13].Jesse Read, Concha Bielza and Pedro Larranaga, 2014. "Multi-dimensional classification with superclassses" in IEEE transactions on knowledge and data engineering, vol. 26, no. 7, july.
[14].Jiawei Han, Micheline Kamber, Jian Pei, 2012. "Data Mining: Concepts and Techniques",Third Edition (The Morgan Kaufmann Series in Data Management Systems).
[15]. M. Hall et al., 2009. "The WEKA data mining software: An update," SIGKDD Explor., vol. 11, no. 1, pp. 10–18.
[16].Margaret H Dunham, 2003. "Data Mining: Introductory and Advanced Topics".
Min-ling zhang, zhi-Hua Zhou, 2014. " A Review on Multi-label Learning Algorithms", in in IEEE transactions on knowledge and data engineering, vol. 26, no. 8, august.
[17].Salvador Garcia and Francisco Herrera, 2008. "An extension on "statistical comparisons of classifiers over Multiple Data sets" for all pairwise comparisons" in JMLR.
[18]. Vincent Labatut, Hocine Cherifi, 2011. "Evaluation of Performance Measures for Classifiers Comparison", ICIT conference.
[19].M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, 2007.
[20]. A. de Carvalho and A. A. Freitas, "A tutorial on multi-label classification techniques," in Studies in Computational Intelligence 205, A. Abraham, A. E. Hassanien, and V. Snásel, Eds. Berlin, Germany: Springer, 2009, pp. 177–195.
[21]. Margaret H Dunham, "Data Mining: Introductory and Advanced Topics", 2003.
[22]. Jiawei Han, Micheline Kamber , Jian Pei, "Data Mining: Concepts and Techniques",Third Edition (The Morgan Kaufmann Series in Data Management Systems), 2012.
[23]. B. Ženko and S. Džeroski, "Learning classification rules for multiple target attributes," in Proc. 12th PAKDD, Osaka, Japan, 2008, pp. 454–465.
[24]. J. H. Zaragoza, E. Sucar, E. F. Morales, C. Bielza, and P. Larrañaga, "Bayesian chain classifiers for multidimensional classification," in Proc. 24th IJCAI, 2011, pp. 2192–2197.
[25]. G. Tsoumakas, I. Katakis, and I. P. Vlahavas, "Mining multilabel data," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Berlin, Germany: Springer, 2010.
[26] G. Tsoumakas and I. P. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in Proc. 18th ECML,Warsaw, Poland, 2007, pp. 406–417.
[27]. J. Read, "Scalable Multi-label Classification," PhD thesis, Univ. Waikato, Hamilton, New Zealand, 2010.
[28] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in Proc. 8th IEEE ICDM, Pisa, Italy, 2008, pp. 995–1000.