非平衡数据集分类方法探讨

职为梅 郭华平 范 明 叶阳东

(郑州大学信息工程学院 郑州 450052)

摘 要 由于数据集中类分布极不平衡,很多分类算法在非平衡数据集上失效,而非平衡数据集中占少数的类在现实生活中通常具有显著意义,因此如何提高非平衡数据集中少数类的分类性能成为近年来研究的热点。详细讨论了非平衡数据集分类问题的本质、影响非平衡数据集分类的因素、非平衡数据集分类通常采用的方法、常用的评估标准以及该问题中存在的问题与挑战。

关键词 非平衡数据集,分类,抽样技术,代价敏感学习

中图法分类号 TP181 **文献标识码** A

Discussion of Classification for Imbalanced Data Sets

ZHI Wei-mei GUO Hua-ping FAN Ming YE Yang-dong

(College of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract Because of imbalanced class distribution, most classifiers lose efficiency with it. In fact the rarely occurring class in imbalanced datasets shows statistical significance. The problem of learning from imbalanced datasets has attracted growing attention in recent years. The paper provided a comprehensive review of the classification of imbalanced datasets, the nature of the problem, the factor which affected the problem, the current assessment metrics used to evaluate learning performance, as well as the opportunities and challenges in the learning from imbalanced data.

Keywords Imbalanced data sets, Classification, Sampling methods, Cost-sensitive learning

1 引言

分类是知识发现和数据挖掘的一项主要任务,分类模型是从训练数据集中学习得到的一个函数,它可以预测未知样本的类标号[1]。现在已经有很多成功的分类算法,比如,判定树、神经网络、贝叶斯网络、支持向量机等,还有很多最近被提出来的算法,他们在许多应用领域都取得了成功。然而,研究证明,这些算法在数据集中各个类分布相对平衡时性能很好,当数据集中各个类分布不平衡时性能很差[2-4]。

不平衡的类分布[2]指的是在一个数据集中一个或一些类实例数很少(也即稀有类),而另一个或一些类实例很多(多数类)。通常把这些数据集的分类问题称为非平衡数据集分类(也称稀有类分类)。普通的分类器在非平衡数据集上往往失效,因为它们建立在训练数据集上并输出最简单的假设适应这些数据,稀有类数据在训练数据集中占很小的比例,分类器倾向多数类数据,因此其在稀有类数据上的效果很差,而稀有类数据往往是有显著意义的数据,比如,在网络侵入检测中,多数连接都属于正常的访问,少数连接数据属于攻击类数据或者黑客数据,有效识别这些数据对网络安全很有意义。因此,近年来,非平衡数据集分类问题成为了数据挖掘的一个热点[2]。

近年来,有一批研究者关注非平衡数据集分类问题,他们提出了很多有效的方法和技术来提高该问题的分类性能。本文探讨了非平衡数据集分类问题的本质,并给出目前解决该问题的一些通用方法;同时对非平衡数据集分类问题进行展望,分析该问题中存在的机遇和挑战。

本文第2节讨论非平衡数据集分类问题的本质;第3节介绍非平衡数据集常用的分类方法;第4节给出通用的衡量标准;第5节分析存在的机遇和挑战;最后总结。

2 非平衡数据集分类的本质

严格地说任何类分布不一致的数据都属于非平衡数据集。然而,业界通常认为类分布极不平衡的数据集才是真正意义上的非平衡数据集,对它的分类称为非平衡数据集分类或者稀有类分类[3]。下面给出的例子是一个非平衡的数据集,通过这个例子了解什么是非平衡数据集分类问题。

考查乳腺癌数据集,该数据集是对病人做乳腺检查得到的。数据集中的每一个记录都有一个类标号"正例"或者"反例",其分别代表乳腺癌患者或者非乳腺癌患者。我们希望反例数目远远超过正例数目,实际也是如此。该数据集包括10923条反例和260条正例。基于该数据集学习分类器,我们希望分类器能够准确对未知样本分类,这意味着能够将一个乳腺癌患者正确识别出来。而实际上由于正例数据很少,分类器往往将多数正例数据认为是反例,也就是说将癌症患

本文受国家自然科学基金项目(60773048)资助。

职为梅(1977-),女,硕士生,讲师,主要研究方向为数据挖掘,E-mail:iewmzhi@zzu.edu.cn;郭华平(1981-),男,博士;范 明(1948-),教授,博士生导师;叶阳东(1962-),教授,博士生导师。

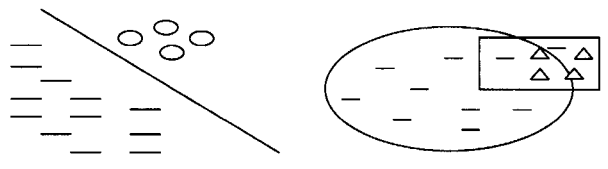
者视为非癌症患者。在实际应用中，这种误分代价远远超过将一个非癌症患者诊断为癌症患者[3]。因此，在这些情况中，需要一个分类器在不损害多数类分类正确率的前提下提高稀有类分类的准确率。当然，准确率并不是衡量稀有类分类的唯一标准，我们将在第4节讨论更多的衡量稀有类分类的标准，这些标准比准确率的效果更好。不仅医学领域存在非平衡问题外，在实际的很多领域都存在这样的问题，比如，网络入侵、石油探测、信用卡欺诈。

表面上看，不平衡的类分布是导致稀有类分类性能差的主要因素，实际上还存在其他影响稀有类分类的因素。文献[5-10]的研究和实验表明，样本的大小、数据的可分离性、类内子概念等因素也极大地影响了稀有类分类。

不平衡的类分布：以包含两个类的数据集为例，类分布指稀有类的实例总数与多数类的实例总数的比值，比如，1:1000。很明显，对于类分布相对平衡的数据集来说，各个类提供给分类器的信息均等，因此分类器对每个类都比较公平。而类分布不平衡的数据集中，稀有类实例数目所占比率很小，使得分类器不能获得足够的信息，从而对稀有类不利。文献[5]探讨了训练数据集的类分布与判定树分类性能的关系，但是不能确定多大的类分布比率会导致分类性能下降。研究表明，在有些应用中1:35的时候不能很好地建立分类器，而有的应用中1:10就很难建立了[8]。

样本的大小：对于同一个非平衡性数据集，可以将其分为相对稀少和绝对稀少[3]，比如，包含了1000条实例的乳腺癌数据集，假设癌症患者类和非癌症患者类的比例为1:100，也就是说，在这个数据集中稀有类的实例数只有10个。在这种情况下，不管使用什么样的方法都很难将少数类样本分类正确，因为没有足够的样本建立模型，称这样的数据集为绝对稀少。假设我们测试更多的病人(100倍)，并且类比例不变，这样少数类样本数目达到1000个，就有足够多的样本建立模型来识别稀有类样本，分类性能会好很多。也就是说，如果样本数据集中数据太少，分类性能就差，对于同样的类分布，样本数据集变大，分类性能就会显著变好[9]。

数据的可分离性：如果一个数据集是高度线性可分的，即使类分布很不平衡，普通的分类器也可以获得好的分类性能，比如图1(a)所示的数据集。如果一个数据集不是线性可分的，比如图1(b)所示，即使是良好设计的分类算法其分类的性能也会很差。文献[9]的实验证明，线性可分的数据集对非平衡性不敏感；随着数据集复杂程度的增加，对非平衡性的敏感程度增加。



(a)线性可分的数据集　　(b) 高度复杂的数据集

图1

类内子概念的存在：非平衡数据集可以分为类间不平衡和类内不平衡[10]。前面提到的数据集均属于类间不平衡，类内不平衡的数据集是指单个类内又有子类或子概念存在，使得数据集进一步变得复杂，从而导致分类性能下降，如图2所示。A、B、C属于一个大类，它们又分别是这个大类下的3个子概念；D和E属于一个类，是原始数据集中的小类，它们又分别属于这个小类中的两个子概念。对于这样的数据集应该采用特殊的方法处理，然后再建立分类器，我们将在3.2.4节对其进行讨论。

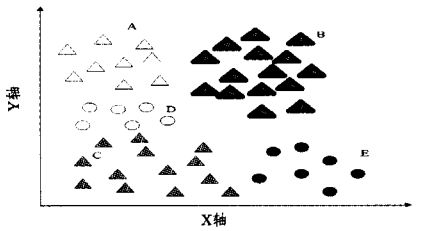


图2　类内存在子概念的数据集

3　非平衡数据集常用分类方法

判定树[6,7]、支持向量集[11-13]、神经网络[9]、贝叶斯网络[14]、最邻近算法[5,15]等传统的分类算法在非平衡数据集上失效，引起大量的专家学者对非平衡数据集分类问题进行了详尽的研究。目前的研究在解决该问题的方法上总体可以分为两类：基于数据的方法和基于算法的方法[2]。基于数据的方法思想如下：通过抽样技术使得不平衡的数据集变得平衡，再在平衡后的数据集上应用传统的分类算法，常用的技术有增加稀有类实例或者减少多数类实例。基于算法的方法思想如下：通过改进已有分类算法使得它更倾向于稀有类，从而提高非平衡数据集的分类性能，例如两阶段归纳方法。

3.1　基于数据的方法

基于数据的方法主要是通过各种抽样技术改变数据集的类分布，使得不平衡数据集变得平衡，从而提高分类性能。抽样技术又可分为随机过抽样(randomly over sampling)、随机欠抽样(randomly under sampling)、有指导的抽样(informatively samping)和综合过抽样(SMOTE)等[3]。

3.1.1　随机过抽样和欠抽样

随机过抽样技术是从稀有类实例集S_{min}随机复制样本集E增加到非平衡数据集S中，使得数据集平衡。这样稀有类样本总数为$|S_{min}|+|E|$，可以调整E的大小得到任意程度的平衡数据集，从而获得好的分类性能。但是过抽样会导致分类器过分拟合训练数据集。

随机欠抽样(under-sampling)是从多数类实例集S_{max}随机删除样本集E。这样总实例数为$|S_{min}|+|S_{max}|-|E|$，同样也可以调整E的大小使得数据集达到任意程度的平衡，从而获得好的分类性能。但是欠抽样会使分类器失去一些包含多类信息的实例，从而导致分类器对多数类不利。

3.1.2　有指导的欠抽样

BalanceCascade方法是一种有指导的欠抽样方法。它的基本思想是从多数类数据S_{max}中取数据集E，使得$|E|=|S_{min}|$，使用$T=\{E\cup S_{min}\}$构成训练数据集学习分类器$C(1)$，观察$C(1)$的结果，标识所有被$C(1)$正确分类的多数类样本x，并从S_{max}中去掉这样的x；然后再从S_{max}取E，且$|E|=|S_{min}|$，使用$T=\{E\cup S_{min}\}$获得$C(2)$。迭代该方法直到获得一组联合的分类器用于分类[16]。

3.1.3　综合抽样

综合过抽样技术(SMOTE)[17]在各种应用中已经表现出了巨大的成功，它的基本思想是考察稀有类样本的特征空间，向数据集中增加人工数据。具体做法如下：给定一个稀有类

实例 $x \in S_{min}$，对于特定的 k 值，计算 x 的 k 近邻，随机选择 k 个近邻中的一个来产生人工数据：

$$x_{new} = x + (x_i - x) \times \delta \quad (1)$$

式中，$x \in S_{min}$，x_i 是 x 的一个近邻，$\delta \in [0,1]$是一个随机数，根据式(1)，x_{new}被增加到数据集中。

综合抽样不是简单地复制稀有类实例，也就不存在随机过抽样存在的过分拟合数据集问题，并且这种方法通过一种特定的方式平衡了原始数据集，从而提高了分类的性能。

3.2 基于算法的方法

基于算法的方法是通过改进现有分类方法，使其更倾向于稀有类数据。在过去的 10 年中，很多研究者提出了改进算法用于非平衡数据集分类。由于篇幅的限制，下面只介绍具有典型意义的改进算法。

3.2.1 组合方法

组合方法[1]又称为集成方法，它的基本思想是由训练数据构建一组基分类器，然后通过对每个基分类器的预测进行投票来分类，也即聚集多个分类器的预测来提高分类准确率。常用的组合方法有装袋(bagging)[19]、提升(Boosting)[20]、随机森林(random forest)[21]等。其中提升是被广泛使用的组合技术。提升是一个迭代的过程，用来自适应地改变训练样本的分布，使得基分类器聚焦在那些很难分的样本上，通过提升，多个弱分类器可以组合成一个强分类器，能有效改善非平衡数据集的分类性能。AdaBoost[22]是提升算法的代表，它对训练集数据的分布迭代加权，在每次迭代中，提升算法增加错误分类的样本权值，减少正确分类的样本权值，而稀有类数据往往被错误分类，这使训练系统在下次迭代中更关注于稀有类样本。因此，该方法对非平衡数据集分类有利。

3.2.2 集成抽样和提升方法

将抽样方法和组合分类器技术集成起来用于非平衡数据集分类的方法也取得了很大的成功，比如 SMOTEBoost[18]。SMOTEBoost 方法是 SMOTE 和 AdaBoost 方法的集成。具体做法是在每次迭代时加入综合抽样技术，这样每个后继的分类器都聚焦于稀有类，由于每个分类器建立在不同的样本数据上，组合分类器联合每个分类器的投票最终决定样本的类别，因此该方法对稀有类分类有效。

3.2.3 两阶段规则归纳方法(PNrule)

两阶段规则归纳方法(PNrule)[23]本质是基于规则的分类方法，采用两个阶段训练规则后用于非平衡数据集的分类，是基于改进算法方法中的一个典型例子。

给定训练数据集，PNrule 算法的基本思想如下：将训练过程分为两个阶段，第一个阶段从整个数据集开始，采用顺序覆盖技术迭代产生规则，得到的规则尽量覆盖多的正例(稀有类实例)，不考虑其覆盖的反例(多数类实例)，该阶段训练得到的规则称为 P 规则，第一个阶段追求高覆盖率，尽可能多地覆盖稀有类数据；第二个阶段从所有的 P 规则覆盖的样本集开始，P 规则覆盖的正例和反例分别变为第二阶段的反例和正例，训练得到的规则称为 N 规则，第二个阶段追求高精确率，尽可能多地删除多数类实例，如图 3 所示。

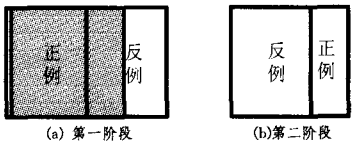


图 3 两阶段分类方法

用第一组规则对未知样本 x 分类，如果分到多数类，则 x 属于多数类；否则，用第二组规则确定 x 所属的类。

在第一个阶段中，可以尽可能地覆盖更多的正例，而不用对其准确度做太多的考虑，因为在第二阶段中训练规则会去除反例。第二阶段的存在使得模型有很好的能力获得稀有类不存在的特性，因为它将所有的反例连接起来。在第二阶段学习 N 规则去除反例，故可以提高规则的精确度。

由上面的分析可以得出前面的结论：两阶段方法能够在两个阶段中分别获得高覆盖和高精确度。两个阶段相结合，从而在稀有类分类上具有很强的优势。实验结果表明，基于规则的两阶段分类法具有较好的分类效果，特别适合对稀有类进行分类，其分类的误差和对稀有类的误分类率都显著低于 C4.5 和 Ripper。

3.2.4 代价敏感学习方法

代价敏感分类技术[24,25]在构建模型的过程中考虑代价矩阵，并产生代价最低的模型。代价矩阵是对一个类的样本分类到另一个类的惩罚。令 $C(i,j)$表示预测一个 i 类记录为 j 类的代价，$C(+,-)$是把正例(稀有类)误分为反例(多数类)的代价，$C(-,+)$是把反例误分为正例的代价。在非平衡数据集分类中，通常认为识别正例的意义大于识别反例的意义，因此，把正例预测为反例的代价高于把反例预测为正例的代价(即 $C(+,-)>C(-,+)$)，而正确分类的惩罚为 0(即$C(+,+)= C(-,-)$)。产生代价最小的模型有利于非平衡数据集分类。

AdaCost[26]是一种基于代价敏感的 boosting 方法。已经证明当以精度和召回率衡量分类结果时，AdaCost 的性能比其他版本的 Boosting 算法都好。

3.2.5 基于局部聚类方法

对于类内存在子概念的非平衡的数据集来说，上述的方法往往都失效。可以采用局部聚类方法[27]对数据集聚类，实验证明这种方法能够获得较好的分类性能。具体做法如下：首先对包含了子类概念的类进行聚类，得到多个小类，在某种程度上使得数据集变得平衡。如图 4 所示，A、B、C 属于一个大类，它们又分别为这个大类下的 3 个子概念；D 和 E 属于一个类，是原始数据集中的小类，它们又分别属于这个小类中的两个子概念。采用局部聚类方法后，A、B、C、D 和 E 被聚类为一个独立的类，从而整个数据集就变得平衡，再对聚类后的数据集使用普通的分类方法(比如支持向量机)训练模型分类。文献[13]的实验结果证明，这种方法在复杂的非平衡数据集上可以得到很好的分类效果。

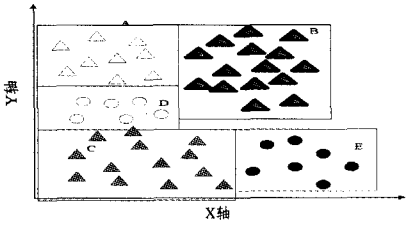


图 4 局部聚类

如果聚类后整个数据集中稀有类实例的数目仍旧很少，还可以采用过抽样技术来增加稀有类实例，使得数据集进一步平衡。

4 稀有类分类的评估标准

准确率是分类中常用的评估标准[2]。对于平衡数据集的

分类来说，准确率的确是一个好的衡量方法。然而对于非平衡数据集分类问题来说，由于关注的焦点不同，仅用准确率是不合适的。

例如：在一个非平衡数据集中，稀有类实例仅占1%，多数类数据占99%。如果一个分类器将所有的未知样本都分为多数类，总体上分类准确率达到99%，这看似相当准确，实际上这个分类器对于稀有类而言性能极差，没有一个稀有类样本被正确分类。因此，在稀有类分类问题中应该采用其他的评价标准。

下面以二元分类为例，讨论其他的分类标准。假设 C 类为稀有类，NC 为多数类，根据分类器的预测类标号和实际类标号的情况，存在如表1所列的混合矩阵。

表1 二元分类问题的混合矩阵

	预测为C类	预测为NC类
实际为C类	TP	FN
实际为NC类	FP	TN

根据表1得到如下度量：

$$Recall=\frac{TP}{TP+FN};Precision=\frac{TP}{TP+FP}$$

Recall 也就是召回率，度量被分类器正确预测的稀有类的比例。具有高召回率的分类器很少将稀有类误分为多数类。

Precision 即精确率，确定在分类器断言为稀有类的那部分实例中稀有类实例实际所占的比例。具有高精度的分类器很少将多数类误分为稀有类。

有的分类器召回率很高但精度很差，有的分类器有很高的精度但召回率很低，这都不是好的分类器。构建一个最大化精度和召回率的模型是稀有类分类的主要任务之一。为了综合考察一个分类器的精度和召回率，可以将精度和召回率可以合并成一个度量，即 F-度量(F-measure)，它的定义如下：

$$F=\frac{2RP}{R+P}$$

式中，R 为 *Recall*，P 为 *Precision*。在非平衡数据集分类中，通常认为 F-度量值越高分类器性能越好。

5 机会与挑战

目前关于非平衡数据集的分类主要集中于算法的研究上，已经提出的算法很少有直接应用在原始数据上的，虽然这些算法都报告称在标准数据集中取得了很好的分类性能，但它们可能过分拟合数据集[3]。因此，研究设计可以处理大容量原始数据集的算法更有意义。

非平衡数据集的研究已经引起了极大的关注，在研究非平衡数据集分类时，下面的问题值得仔细思考：

1)原始数据集应该被平衡到什么程度算最好？

2)在对非平衡数据集分类时，应该做怎样的假设才能取得和平衡数据集分类同样好的效果？

3)非平衡的类分布和分类算法的计算复杂度之间有什么样的关系？也就是说我们将来的研究重点放在数据的理解上，如果能够做好这一步，也许数据集的非平衡性就不是分类的主要问题了。

4)特征选择也是非平衡数据集分类的一个重要选择，考虑在分类前选择合适的特征，是否也可以改善分类性能？

结束语 本文是对非平衡数据集分类问题研究的总结，分析了非平衡数据集分类问题的本质、非平衡数据集分类的评估标准、目前流行的分类方法(基于数据的方法和基于算法改进的方法)以及介绍了这些方法的基本思想。希望这些能够为想要从事非平衡数据集分类研究的学者提供一些指导。

参考文献

[1] Tan Pang-ning, Steinbach M. Introduction to Data Mining(第2版)[M]. 范明，范宏建，译. 北京：人民邮电出版社，2011：127-187

[2] Sun Yan-min, Kamel M S, Wong A K C. Cost-sensitive boosting for classification of imbalanced data. Patter Recognition Society [J]. Published by Elsevier Ltd, 2007：3358-3378

[3] He Hai-bo, Garcia E A. Learning from imbalanced Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9)：1263-1284

[4] Visa S, Ralescu A. Issues in Mining imbalanced Data Sets-A Review Paper[C]// Proc. of MidWest Artificial Intelligence and Cognitive Science Conference (MAICS'05). Dayon, 2005：67-73

[5] Batista G E A P A, Prati R C, Monard M C. A study of the Behavior of several methods for balancing machine learning training data[J]. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 2004, 6(1)：20-29

[6] Japkowicz N, Stepen S. The class imbalance problem：a systematic study[J]. Intell. Data Anal. J., 2002, 6(5)：429-450

[7] Weiss G, Provost F. Learning when training data are costly：the effect of class distribution on tree induction[J]. J. Aritif. Intell. Res., 2003, 19：315-354

[8] Joshi M V. Learning classifier models for predicting rare phenomena[D]. University of Minnesota, Twin Cites, MN, USA, 2002

[9] Japkowicz N, Stephen S. The class imbalance problem：a systematic study[J]. Intell. Data Anal. J., 2002, 6(5)：429-450

[10] Japkowicz N. Concept-learning in the presence of between-class and within-class imbalance[C]// Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence. Ottawa, Canada, June 2001：67-77

[11] Akbani R, Kwek S, Jakowicz N. Applying support vector machines to imbalanced datasets[C]// Proceedings of European Conference on Machine Learning. Pisa, Italy, September 2004：39-50

[12] Raskutti B, Kowalczyk A. Extreme rebalancing for SVMs：a case study[C]// Proceedings of European Conference on Machine Learning. Pisa, Italy, September 2004：60-69

[13] Wu G, Chang E Y. Class-boundary alignment for imbalanced dataset learning[C]// Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. Washington, DC, August 2003

[14] Ezawa K, Singh M, Norton S W. Learning goal oriented Bayesian networks for telecommunications risk management[C]// Proceedings of the Thirteenth International Conference on Machine Learning. Bari, Italy, 1996：139-147

[15] Zhang J, Mani I. KNN approach to unbalanced data distributions：a case study involving information extraction[C]// Proceedings of the ICML'03 Workshop on learning from Imbalanced Data Sets. Washing, DC, August 2003

[16] Liu Xu-ying, Wu Jian-xin, Zhou Zhi-hua. Exploratory Under

Sampling for Class Imbalance Learning[C]// Proc. Int'L Conf. Data Mining. 2006:965-969

[17] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-Sampling Technique[J]. J. Artificial Intelligence Research, 2002, 16:321-357

[18] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting[C]// Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases. 2003:107-119

[19] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2):123-140

[20] 沈学华,周志华,吴建鑫. Boosting 和 Bagging 综述[J]. 计算机工程及应用,2000,36(12)

[21] Breiman. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32

[22] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]// Proceedings of the Thirteenth International Conference on Machine Learning. The Mit Press, Cambridge, MA, Morgan Kaufmann, Los Altos, CA, 1996:148-156

[23] Agarwal R, Joshi M V. PNrule: A new Framework for Learning Classifier Models in Data Mining (A Case-Study in Network Intrusion Detection)[R]. Department of Computer Science University of Minnesota, USA, 2000

[24] Elkan C. The Foundations of Cost-Sensitive Learning [C]// Proc. Int'l Joint Conf. Artificial Intelligence. 2001:973-978

[25] Ting K M. An Instance-Weighting Method to Induce Cost-sensitive Trees[J]. IEEE Trans. Knowledge and Data Eng., 2002, 14(3):659-665

[26] Wei Fan, Stolfo S, Zhang Jun-xin. AdaCost: Misclassification Cost-sentitive Boosting[C]// Proceedings of the 16th International Conference on Machine Learning. 1999:97-105

[27] Wu Jun-jie, Xiong Hui, Chen jian. COG: local decomposition for rare class analysis[J]. DMKD, 2010, 20(2):1384-5810

(上接第 292 页)

图 4 展示了动态分析的界面。

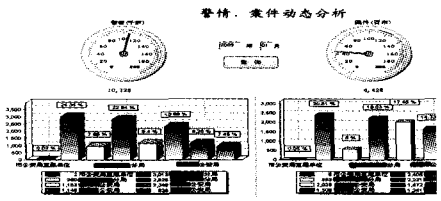


图 4 动态分析图

5.2 主题分析

主题分析主要是用来针对特殊的或者阶段性的分析需求而专门进行的分析作业。图 5 示出了案件主题的分析。

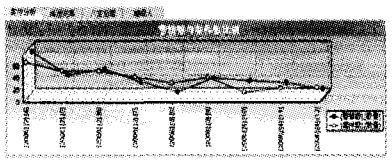


图 5 主题分析图

5.3 情报研判

情报研判管理主要针对捕获数据的汇总、分析。用户可结合信息查询、全文检索、BI、i2 等工具对"比对结果分析"做更深层次的分析,最终得出数据的研判评估。并将研判评估的结果发布到应用平台门户网站上。

另外,还提供了信息查询、统计报表等功能。

结束语 本文详细介绍了在公安行业建设与实施数据仓库系统的问题以及解决的方法,给出一个适合公安行业数据仓库的可行性解决方案。ETL 方案为数据仓库的建设与实施建立了良好的数据基础,保证了数据仓库中数据的正确性、完整性。合理地设计了系统的人、案件、物品、机构、地点等 5 大主题模型,决策层可以根据每个主题的展示结果来分析数据,掌握各个地区的各类案件的发案、立案、破案、涉案物品、涉案人员等 5 大主题的情况。

基于此数据仓库,实现了警务综合信息系统的数据共享、综合查询、统计分析、自定义报表、情报导侦与外部职能部门的数据交换、决策支持分析等各种数据应用。通过对系统中各大主题分析结果进行相互关联分析,来更好地预防犯罪,提高破案率,提高"打"、"防"、"管"、"控"的工作效率。

本文相关研究成果已在实际应用中得到验证及完善,并在实际的应用过程中起到了较好的效果。2011 年某月份全省社会警情通报最新数据显示,杀人、绑架、强奸案件同比下降 2%、24.1%、7.8%,入室抢劫、抢汽车案件同比下降 24.8%、51.2%,放火、投放危险物质案件同比下降 8.7%、66.7%,涉枪案件同比下降 12.6%。

通过本课题的研究,得到一个适合公安行业数据仓库的可行性解决方案,并与"金盾工程"建设相结合,把公安情报信息资源作为新时期公安工作的重要战略资源。

参 考 文 献

[1] 王春雨,王延章,叶鑫,等. 基于数据仓库的刑事案件决策支持系统设计[J]. 计算机工程与设计,2010,31(4):767-775

[2] 麦永浩,杨超. 公安数据仓库和数据挖掘应用研究[J]. 警察技术,2009(2):27-29

[3] 尉宁. 电信行业数据仓库建设与实践[D]. 重庆:重庆大学,2007

[4] 徐珊. 数据仓库在江西省公安厅综合信息分析系统中的应用[D]. 江西:南昌大学,2008

[5] 杨兴凯. 基于本体的政务数据仓库构建方法研究[J]. 计算机工程与设计,2010,31(7):1492-1499

[6] 汪涛. 医院数据仓库数据模型设计[J]. 计算机应用与软件,2010,5:191-194

[7] Joy Mundy Warren Thornthwaite Ralph Kimball. 数据仓库工具箱[M]. 北京:清华大学出版社,2007:1-265

[8] Golfarelli M, Rizzi S. 数据仓库设计——现代原理与方法[M]. 战晓苏,吴云浩,皮人杰,译. 北京:清华大学出版社,2010:1-375

[9] 英蒙. DW2.0 下一代数据仓库的架构[M]. 北京:机械工业出版社,2010:1-198

[10] 周亮. 电子政务决策支持系统中数据仓库的研究与设计[J]. 武汉理工大学学报,2005,27(1):31-35